1996

Statistical analysis of maintenance growth curves

Thomas James Kirchoff
Iowa State University

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Statistical analysis of maintenance growth curves

by

Thomas James Kirchoff

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Major Professor: Dr. Kenneth J. Koehler

Iowa State University

Ames, Iowa

1996

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Thomas James Kirchoff

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my sincere thanks: To Dr. Kenneth Koehler, for his support and guidance in the direction of this dissertation and for his helpful advice on my professional career; to Dr. Robert Stephenson who has served as my mentor for teaching throughout my stay at Iowa State; to Dr. David Cox for serving as my master's advisor; to Dr. David Cox and Dr. Noel Cressie for both encouraging me to continue the doctoral program at a time when I considered otherwise; to Dr. James Cornette of the mathematics department for serving on the committee; to Dr. H.T. David for serving on the committee on rather short notice; to my wife Amy and my children Kathryn, Paul, and Mary for their patience and understanding.

1 OVERVIEW

Introduction

Longitudinal data analysis is the analysis of data on subjects which is collected over time. The repeated observations on subjects tend to be correlated and this correlation must be taken into account for proper inferences to be made.

Longitudinal data analysis is often referred to as growth curve analysis when attention is centered on the time response curve for the population or for an individual subject. Classical growth curve analysis is based on balanced designs where every subject is measured at the same time points with no missing observations. In this case multivariate analysis methods which assume a general covariance structure for observations taken on the same subject can be used. Rao [21] introduced the idea of random parameters and the idea was developed further in the context of growth curves by Rao [22], Grizzle and Allen [10], and Potthoff and Roy [19]. Models for growth curves with unbalanced data, where each subject is observed at a different set of time points, were introduced by Laird and Ware [17] using the work of Harville [13].

Often the problem of interest in growth curve analysis is the prediction of a future observation for a subject given the observations at previous time points. Rao [24] considered this problem for growth curves in the case of balanced data with no missing observations. In this thesis we consider the problem of modeling an average growth curve in the case of unbalanced data using both parametric and nonparametric methods. We will also be interested in the prediction of the increase of the average growth

curve for a given time interval as well as the prediction of a future observation for a subject given the past observations.

As an example of data which can be analyzed as growth curves we consider data from a large truck leasing firm located in the midwestern United States. This firm owns a large fleet of trucks which are in turn leased to several different companies throughout the country. As part of the leasing agreement with these companies the firm agrees to perform periodic maintenance and repairs on the trucks. The maintenance and repairs are carried out in one of several maintenance centers located throughout the country which are managed and staffed by the firm. The firm is interested in the number of labor hours expended by its staff in the maintenance and repair of these trucks. Information on the number of labor hours needed to maintain the trucks is useful for forecasting staffing needs at the various maintenance centers.

The data

The firm records labor hours incurred for each truck in an annual report at the end of the service year. The trucks in the study were put into service between the years 1987 and 1993. Between one and seven annual reports are available for a truck depending on the year it was placed into service.

The trucks in the fleet are large tractor units designed to pull trailers. All trucks have been classified into six categories for record keeping purposes based on configuration. Features which determine a truck's configuration are the number of rear axles (one or two), engine placement (cabover or conventional) and the presence or absence of a sleeper. The largest category of trucks had 1182 units and it was data from this category that were analyzed for this thesis. The total number of annual reports in the largest category was 5344. An annual report for a truck consists of cumulative labor hours expended on the truck to date and cumulative mileage on the truck. A scatterplot of cumulative labor hours versus cumulative mileage in 1993 is shown in Figure 1.1.

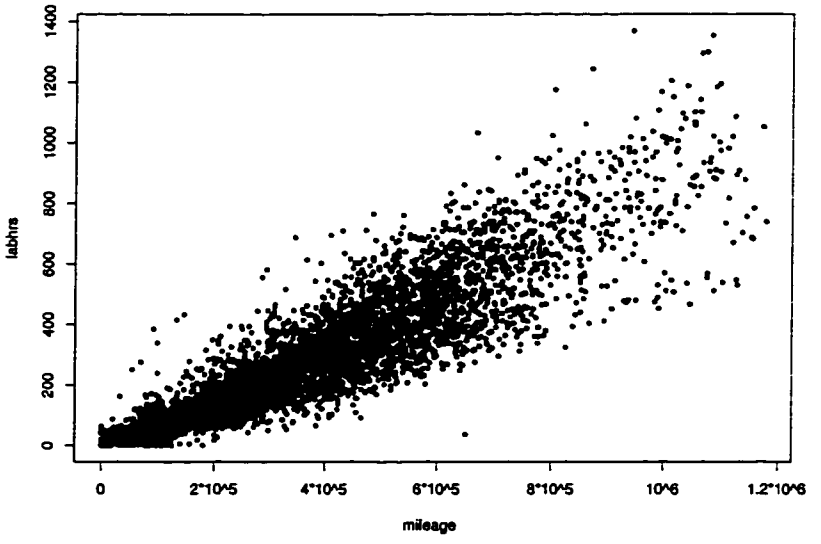

Figure 1.1 Labor hour histories - 1993

Methods used

Letting y =cumulative labor hours and t =cumulative mileage, we can reasonably assume that observations within a truck are correlated and observations between trucks are independent. Thus, the truck data can be viewed as a series of growth curves. Cumulative mileage is used as the time variable rather than age of the truck because it is a measure of both the age and the amount of wear on a truck. Even though the data are collected annually, using cumulative mileage as the time variable implies the data will be unbalanced since each truck is observed at a different set of time points.

We will consider the problem of predicting labor hours needed to maintain the trucks for a future year. In Chapters 2 and 3 we consider nonparametric methods of estimating the average growth curve for the population. Chapter 2 introduces a linear interpolation method to estimate the average growth curve. Locally weighted regression, or *loess*, is used in Chapter 3. Estimated increases in labor hours within fixed mileage intervals can be computed using the average growth curve. Predictions for the labor hours needed to maintain a truck may be obtained by estimating the number of miles the truck will be

driven in the future year.

Chapter 4 considers the use of linear mixed models in the problem of predicting labor hours needed. With linear mixed models it is possible to predict the growth curve for an individual truck as well as estimate the average growth curve for the population. Prediction of the labor hours needed by an individual truck in a future year may be obtained by that truck's growth curve.

2 AN INTERPOLATION METHOD

Introduction

In this chapter we use a nonparametric linear interpolation method to analyze growth curve data obtained from irregular inspection schedules that vary across individual units or respondents. This method is used to estimate values on the mean population growth curve without specifying a functional form for the curve. It is also used to estimate increases in the mean growth curve between any pair of time points. This method has the advantages that estimates of increases in the mean growth curve are always non-negative and they are easily computed.

The interpolation method is applied in the following general situation. Data are obtained on p units and the i^{th} unit is measured at n_i time points. A very general model for describing the relationship between the measured characteristic (Y) and some measure of time or usage (T) is

$$Y_{i,j} = m(T_{i,j}) + \epsilon_{i,j} \qquad \begin{array}{l} i = 1, ..., p \\ j = 1, ..., n_i \end{array} \tag{2.1}$$

where m is an unknown monotone nondecreasing function and $\epsilon_{i,j}$ are observation errors. The observation errors are assumed to have $E(\epsilon_{i,j}) = 0$, $Var(\epsilon_{i,j}) = \sigma^2(T_{i,j}) < \infty$, and $Cov(\epsilon_{i,j}, \epsilon_{i,k}) = \sigma(T_{i,j})\sigma(T_{i,k})\rho(T_{i,j} - T_{i,k})$, $j > k$, where ρ is a correlation function that is even with $\rho(0) = 1$ and $|\rho(y)| \leq 1$. The variance function $\sigma^2(T_{i,j})$ allows for nonconstant variance. This model also allows for the possibility that observations made

on the same unit are correlated, but observations made on different units are assumed to be uncorrelated.

In applying this model to the truck data, we let $y_{i,j}$ denote cumulative labor hours and $t_{i,j}$ denote cumulative mileage for the i^{th} truck at the end of the j^{th} year of service. The function $m(t)$ represents the population mean for cumulative labor hours at cumulative mileage t. For the j^{th} truck, $\epsilon_{i,j}$ represents the deviation in cumulative labor hours from the population mean at cumulative mileage $t_{i,j}$.

Literature review

There are a variety of nonparametric methods that could be used to estimate the overall mean curve, $m(t)$. Most of these methods have been developed for data where each observation is provided by a different individual or experimental unit and the errors are assumed to be independent and identically distributed. Research in the case of correlated errors has primarily focused on kernel based regression introduced by Priestley and Chao [20]. Kernel based regression requires the user to select a smoothing parameter called the bandwidth. The choice of bandwidth is crucial since it affects the smoothness, bias, and variability of the estimate.

Altman [1] and Hart [11] consider bandwidth selection for kernel regression in the presence of serial correlation using the model

$$y_i = m(t) + \epsilon_i \qquad i = 1, ..., n.$$

However, the data are assumed to be from a single unit or process and the methods do not directly apply to the truck data. Clark [3] considered the estimation of $m(t)$ by convolving a linear interpolation between two successive observations with a kernel function, but this method also does not allow for correlated errors arising from repeated measurements from multiple units. Gasser, Muller, Kohler, Molinari, and Prader [9] considered the use of kernel estimators in the analysis of growth curves, but their pro-

cedure applies to the estimation of a growth curve for a single individual and assumes that errors are independent and identically distributed. Hart and Wehrly [12] consider a repeated measurements model of the form:

$$\begin{aligned} y_{i,j} &= m(t_j) + \epsilon_{i,j} \\ i &= 1, ..., m \\ j &= 1, ..., n \end{aligned}$$

where the t_j's are fixed inspection points with $0 \leq t_1 < t_2 < \cdots < t_n \leq 1$, and the $\epsilon_{i,j}$'s are zero mean random variables satisfying

$$\begin{aligned} Cov(\epsilon_{i,j}, \epsilon_{k,l}) &= \sigma^2 \rho(t_j - t_l) \quad i = k \\ &= 0 \quad i \neq k \end{aligned}$$

and ρ is a correlation function depending only on the distance between t_j and t_l. This model assumes observations are available on all m units at each of the j inspection points, and defining $\bar{y}(t_j) = \sum_{i=1}^{m} y_i(t_j)/m$, the model is reexpressed as:

$$\bar{y}(t_j) = m(t_j) + \bar{\epsilon}(t_j).$$

They focus on the proper choice of bandwidth for fitting a kernel based curve through the points $\{(t_j, \bar{y}_{.j})\}$. This method cannot be applied to data like the truck maintenance study where inspection schedules vary in an irregular manner from unit to unit. Even if the trucks *were* inspected on the same schedule, it is unlikely that there would be more than six or seven total time points for each truck. In that case, Hart and Wehrly's method reduces to fitting a nonparametric curve through six or seven points and one might do better by simply fitting a low order polynomial.

Nonparametric linear interpolation

Nonparametric linear interpolation consists of connecting each successive pair of observations on a truck with a straight line. Then an estimate of the mean cumulative labor hours at some point t_0 is obtained using the appropriate line segment for each truck to interpolate at t_0 and averaging the interpolated values.

To estimate $m(t_a)$, the value of the mean growth curve at some fixed cumulative mileage value t_a, first identify all trucks that were driven at least t_a miles. Suppose there are q_a trucks that satisfy that criterion. For the i^{th} such truck, the interpolated value at t_a is

$$\hat{m}_i(t_a) = \frac{(y_{i,j_a+1} - y_{i,j_a})(t_a - t_{i,j_a})}{t_{i,j_a+1} - t_{i,j_a}} + y_{i,j_a} \tag{2.2}$$

where t_{i,j_a} and t_{i,j_a+1} are the cumulative mileage values closest to t_a that satisfy the condition $t_{i,j_a} \leq t_a \leq t_{i,j_a+1}$, and j_a and j_{a+1} indicate the years in which these observations were taken. Finally, $m(t_a)$ is estimated by averaging these interpolated values across the q_a available trucks, i.e.,

$$\hat{m}(t_a) = \frac{1}{q_a} \sum_{i=1}^{q_a} \frac{(y_{i,j_a+1} - y_{i,j_a})(t_a - t_{i,j_a})}{t_{i,j_a+1} - t_{i,j_a}} + y_{i,j_a}. \tag{2.3}$$

One attractive feature of $\hat{m}(t_a)$ is its computational simplicity. Unlike other nonparametric procedures, there is no choice to be made about a bandwidth or the number of nearest neighbors to be included in the estimate. Like kernel regression estimation and k-nearest neighbor regression, however, $\hat{m}$ is not constrained to be monotone nondecreasing.

A graph of $\hat{m}$ is shown in Figure 2.1. Note that the behavior of $\hat{m}$ becomes erratic at the right end of the graph. This is because older trucks tend to be removed from service leaving fewer trucks that enter into the computation of $\hat{m}$ at larger mileage values. Out of 1182 trucks in the study only 57 had cumulative mileages in excess of 1,000,000 miles. Trucks which are still in service after 1,000,000 cumulative miles are primarily driven

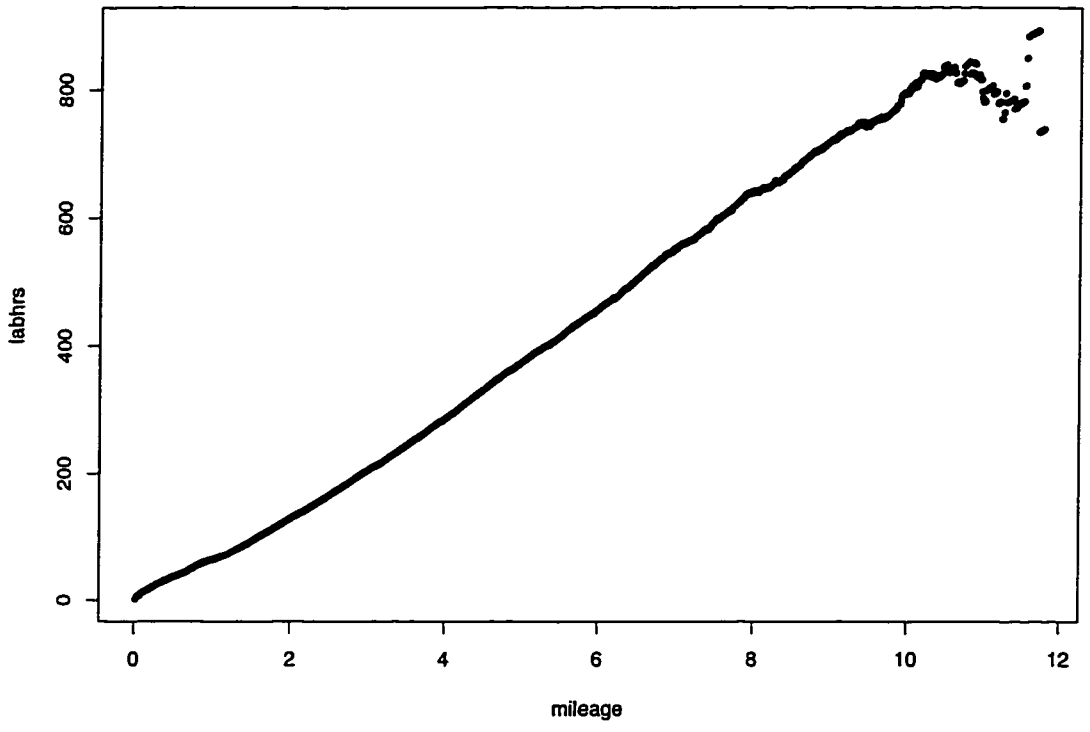

Figure 2.1 Interpolation method

on the highway with high annual mileage rates. These trucks are subject to less wear than trucks with low annual mileage rates which are driven on shorter routes or within cities, and so tend to accumulate fewer labor hours. Thus, as t increases, decreases in $\hat{m}$ can occur. A modification of the interpolation method which constrains predictions to be nonnegative is presented in the next section.

Predictions of cumulative labor hour values for individual trucks can be improved by stratifying trucks according to annual mileage rates and estimating a different mean curve for each stratum. The annual mileage rate, also called the pace, is the average number of miles per year the truck is driven. Aside from the first and last years of operation, which may be only partial years, annual mileage rates on an individual truck are relatively stable across the life of the truck since most customers are companies that tend to sign long term leases and a particular truck tends to be used in a similar manner

from one year to the next. Pace is an indicator of how a truck is used. Low pace trucks are often used to make deliveries within metropolitan areas or for making short hauls between communities. This may increase wear on components such as brakes, engines, and transmissions due to more frequent stops and more engine idling time. High pace trucks that tend to be driven on longer hauls on interstate divided highways may be operated with lower maintenance costs and fewer labor hours.

As many as ten pace categories have been used with good results but we will consider only four pace categories in this thesis to make presentations of graphs and tables more manageable. The four pace categories are identified in Table 2.1. Graphs of $\hat{m}$ for the four pace categories are shown in Figure 2.2. Note that the rate of increase in $\hat{m}$ with respect to t becomes smaller as pace increases, indicating that low pace trucks tend to more rapidly accumulate more labor hours than high pace trucks.

Table 2.1 Four pace categories

Category	Miles per Year
1	0 - 75,000
2	75,000 - 125,000
3	125,000 - 175,000
4	175,000 or more

Forecasting the increase in cumulative labor hours for mileage intervals

We now develop a method for forecasting the *increase* in cumulative labor hours during a given mileage interval, say (t_a, t_b). These forecasts are used to predict future staffing needs at the maintenance centers. Typically, t_a is the cumulative mileage at the end of the current year and t_b is the projected cumulative mileage at the end of the subsequent year.

The estimated increase in labor hours for (t_a, t_b) is given by:

$$\Delta\hat{y}(t_a, t_b) = \hat{m}(t_b) - \hat{m}(t_a). \tag{2.4}$$

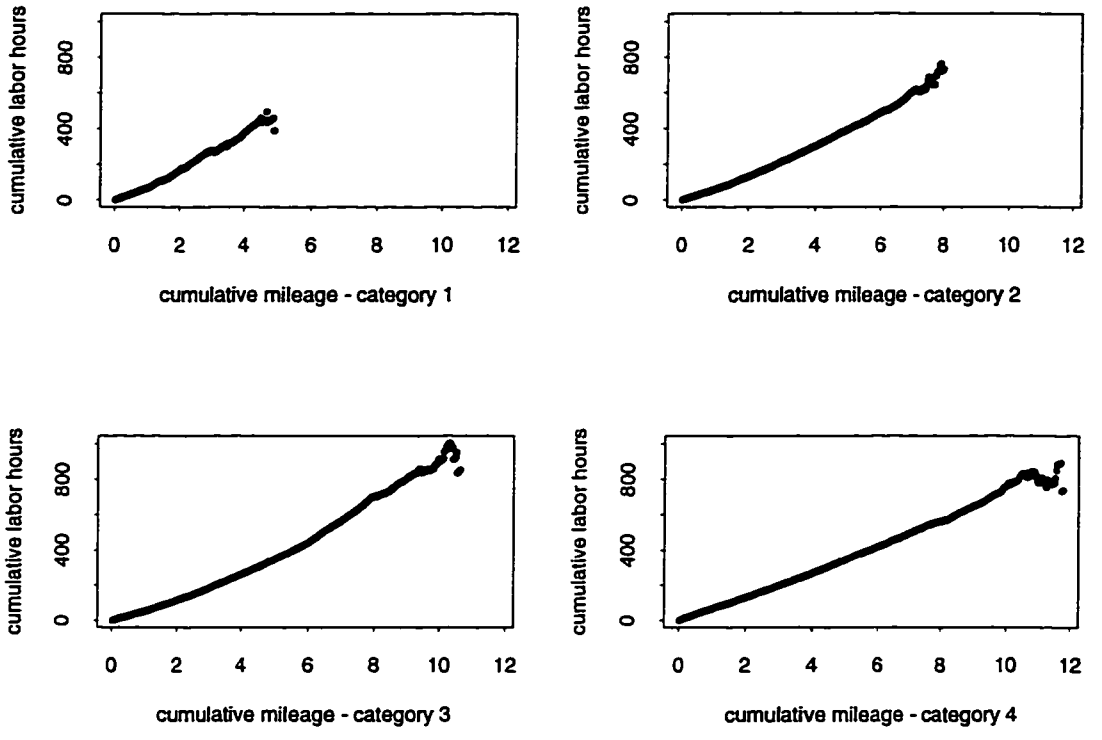

Figure 2.2 Interpolation method - Four pace categories

To constrain $\Delta\hat{y}$ to be positive, $\hat{m}(t_a)$ is calculated only for the q_b trucks which enter into the calculation for $\hat{m}(t_b)$. In other words, $\Delta\hat{y}$ is calculated using only those q_b trucks for which $t_b < T_{i,n_i}$. A disadvantage of this method is that some information is lost when trucks for which $T_{i,n_i} < t_b$ are discarded from the computation. This will result in an increase in $Var(\Delta\hat{y})$, but this method will give $\Delta\hat{y} \geq 0$ for any mileage interval $[t_a, t_b]$.

Table 2.2 gives estimated increases in cumulative labor hours for 50,000 mile intervals for all the trucks and for each of the four pace categories. Within a given mileage interval, the estimated increase in labors hours tends to decrease as pace increases. This result is expected after examining Figure 2.2.

Table 2.2 Estimated increase in labor hours

	Pace category				
Mileage interval	1	2	3	4	All trucks
0-50,000	34.82	30.10	26.20	34.17	30.40
50,000-100,000	33.25	30.09	25.72	32.29	29.67
100,000-150,000	38.25	34.53	29.76	31.05	33.33
150,000-200,000	45.64	38.64	32.66	31.84	36.73
200,000-250,000	55.09	41.24	34.39	33.11	39.19
250,000-300,000	63.79	34.47	36.16	34.03	41.18
300,000-350,000	62.61	45.15	37.97	34.63	42.34
350,000-400,000	74.08	47.80	39.42	35.70	44.28
400,000-450,000	76.68	49.73	41.64	37.05	45.99
450,000-500,000		52.50	44.45	37.51	48.00
500,000-550,000		51.98	44.16	38.00	47.22
550,000-600,000		52.50	45.95	37.54	47.18
600,000-650,000		56.96	49.73	38.25	49.23
650,000-700,000		58.93	51.87	39.08	49.39
700,000-750,000		58.00	51.70	40.00	47.67
750,000-800,000			53.83	39.11	47.21
800,000-850,000			53.03	40.27	46.82
850,000-900,000			57.40	41.82	48.80
900,000-950,000			56.93	42.86	47.23
950,000-1,000,000			52.72	43.77	45.66
1,000,000-1,050,000			62.18	47.88	48.65
1,050,000-1,100,000				42.12	42.12
1,100,000-1,150,000				33.08	33.08

Bias of $\hat{m}(t)$

Unless the true mean growth curve is a straight line, $\hat{m}(t)$ will usually produce biased estimates. The amount of bias depends on the form of the true mean growth curve and the complete set of values T at which observations are taken. In this case T is the entire set of annual cumulative mileage values for all of the trucks in the data file.

The conditional bias of $\hat{m}(t)$, given T, is

$$E(\hat{m}(t)|T) - m(t). \tag{2.5}$$

Assuming $m(t)$ has continuous derivatives through the third order, an approximation to the bias at $t = t_a$ is obtained from a Taylor expansion of $m(t_{i,j_a})$ about $m(t_a)$ and

$m(t_{i,j_{a+1}})$ about $m(t_a)$. We have

$$\begin{aligned} E(\hat{m}(t_a)|T) - m(t_a) &= m(t_a) + \frac{m''(t_a)}{2q_a}\sum_{i=1}^{q_a}(t_a - t_{i,j_a})(t_{i,j_{a+1}} - t_a) + \\ &\sum_{i=1}^{q_a} m'''(z_{i,j_a})\frac{(t_{i,j_a} - t_a)^3(t_{i,j_{a+1}} - t_a)}{6(t_{i,j_{a+1}} - t_{i,j_a})} + \\ &\sum_{i=1}^{q_a} m'''(z_{i,j_{a+1}})\frac{(t_{i,j_{a+1}} - t_a)^3(t_a - t_{i,j_a})}{6(t_{i,j_{a+1}} - t_{i,j_a})} \end{aligned} \tag{2.6}$$

for some $z_{i,j_a} \in [t_{i,j_a}, t_a]$ and some $z_{i,j_{a+1}} \in [t_a, t_{i,j_{a+1}}]$.

If $m(t)$ is approximately quadratic in an interval centered at t_a, then $m'''(t_a) \simeq 0$ and

$$E(\hat{m}(t_a)|T) - m(t_a) \simeq \frac{m''(t_a)}{2q_a}\sum_{i=1}^{q_a}(t_a - t_{i,j_a})(t_{i,j_{a+1}} - t_a). \tag{2.7}$$

We note the following observations about this approximation to the conditional bias:

1. Bias decreases as $m''(t_a)$ decreases. If the true mean curve is a straight line then the bias is 0.

2. Observe $\sum_{i=1}^{q_a}(t_a - t_{i,j_a})(t_{i,j_{a+1}} - t_a) > 0$. Thus, if $m(t)$ is convex in the region $\left\{\min_i t_{i,j_a} < t < \max_i t_{i,j_{a+1}}\right\}$, then the bias approximation (2.7) will be positive at t_a. If $m(t)$ is concave in that region, then the bias approximation is negative.

3. The bias is not reduced as the number of units in the sample tends to infinity.

4. The bias tends to zero as $\max_i(t_{i,j_{a+1}} - t_{i,j_a})$ tends to zero. In other words, bias is reduced if the distance between successive inspection times tends toward zero. Increasing the frequency of inspection times is one way to reduce bias if the true mean curve deviates strongly from a straight line.

For a given mileage interval (t_a, t_b), the approximate conditional bias of $\Delta\hat{y}(t_a, t_b)$ is given by:

$$E(\Delta\hat{y}|T) - [\Delta y] \simeq \frac{m''(t_b)}{2q_b}\sum_{i=1}^{q_b}(t_b - t_{i,j_b})(t_{i,j_b+1} - t_b) - \left[\frac{m''(t_a)}{2q_b}\sum_{i=1}^{q_b}(t_a - t_{i,j_a})(t_{i,j_a+1} - t_a)\right]. \tag{2.8}$$

Variance of $\hat{m}(t)$

The conditional variance of $\hat{m}(t)$ at $t = t_a$ given T is

$$Var(\hat{m}(t_a)|T) = \frac{1}{q_a^2}\sum_{i=1}^{q_a}\left[\frac{(t_a - t_{i,j_a})^2\sigma^2(t_{i,j_a+1}) + (t_{i,j_a+1} - t_a)^2\sigma^2(t_{i,j_a})}{(t_{i,j_a+1} - t_{i,j_a})^2} - \frac{2(t_a - t_{i,j_a})(t_{i,j_a+1} - t_a)\sigma(t_{i,j_a})\sigma(t_{i,j_a+1})\rho(t_{i,j_a+1} - t_{i,j_a})}{(t_{i,j_a+1} - t_{i,j_a})^2}\right]. \tag{2.9}$$

Since m is assumed to have a bounded second moment, we have

$$Var(\hat{m}(t_a)|T) < \frac{1}{q_a^2}[q_a B] \tag{2.10}$$

where $B < \infty$ and thus $\lim_{q_a\to\infty} Var(\hat{m}(t_a)|T) \to 0$.

In equation 2.9 we see that if the $\epsilon_{i,j}$ are positively correlated within a truck, i.e. $\rho(T_{i,j} - T_{i,k}) > 0$, then the variance is smaller than the case where the $\epsilon_{i,j}$ are independent within a truck.

To get an estimate of $Var(\hat{m}(t_a)|T)$ we note that:

$$\begin{aligned} Var(\hat{m}(t_a)|T) &= Var\left(\frac{1}{q_a}\sum_{i=1}^{q_a}\hat{m}_i(t_a)\middle| T\right) \\ &= \frac{1}{q_a^2}\sum_{i=1}^{q_a} Var\left(\hat{m}_i(t_a)|T\right). \end{aligned}$$

A consistent estimator of the variance is simply the sample variance of the independent estimates, $\hat{m}_i(t_a)$. So,

$$\widehat{Var}\left(\hat{m}(t_a)|T\right) = \frac{1}{q_a - 1}\sum_{i=1}^{q_a}\left(\hat{m}_i(t_a) - \hat{m}(t_a)\right)^2. \tag{2.11}$$

For a given mileage interval (t_a, t_b), we have:

$$
\begin{aligned}
&Var\left[\Delta\hat{y}(t_a,t_b)|T\right] = \\
&\quad Var\left[\frac{1}{q_b}\sum_{i=1}^{q_b}\left(y_{i,j_{a+1}}\left(\frac{t_a - t_{i,j_a}}{t_{i,j_{a+1}} - t_{i,j_a}}\right) + y_{i,j_{a+1}}\left(\frac{t_{i,j_{a+1}} - t_a}{t_{i,j_{a+1}} - t_{i,j_a}}\right)\right.\right. \\
&\quad \left.\left. - y_{i,j_{b+1}}\left(\frac{t_b - t_{i,j_b}}{t_{i,j_{b+1}} - t_{i,j_b}}\right) - y_{i,j_b}\left(\frac{t_{i,j_{b+1}} - t_b}{t_{i,j_{b+1}} - t_{i,j_b}}\right)\right) \mid T\right]
\end{aligned}
$$

If we let:

$$
\begin{aligned}
\lambda_{1,i} &= \left(\frac{t_a - t_{i,j_a}}{t_{i,j_{a+1}} - t_{i,j_a}}\right) \\
\lambda_{2,i} &= \left(\frac{t_{i,j_{a+1}} - t_a}{t_{i,j_{a+1}} - t_{i,j_a}}\right) \\
\lambda_{3,i} &= \left(\frac{t_b - t_{i,j_b}}{t_{i,j_{b+1}} - t_{i,j_b}}\right) \\
\lambda_{4,i} &= \left(\frac{t_{i,j_{b+1}} - t_b}{t_{i,j_{b+1}} - t_{i,j_b}}\right) \\
\lambda_i' &= (\lambda_{1,i}, \lambda_{2,i}, \lambda_{3,i}, \lambda_{4,i}) \\
R_i &= \begin{bmatrix} 1 & \rho(t_{i,j_{a+1}} - t_{i,j_a}) & \rho(t_{i,j_b} - t_{i,j_a}) & \rho(t_{i,j_{b+1}} - t_{i,j_a}) \\ & 1 & \rho(t_{i,j_b} - t_{i,j_{a+1}}) & \rho(t_{i,j_{b+1}} - t_{i,j_{a+1}}) \\ & & 1 & \rho(t_{i,j_{b+1}} - t_{i,j_b}) \\ & & & 1 \end{bmatrix} \\
\Delta_i &= \begin{bmatrix} \sigma^2(t_{i,j_a}) & & & \\ & \sigma^2(t_{i,j_{a+1}}) & & \\ & & \sigma^2(t_{i,j_b}) & \\ & & & \sigma^2(t_{i,j_{b+1}}) \end{bmatrix} \\
V_i &= \Delta_i^{\frac{1}{2}} R_i \Delta_i^{\frac{1}{2}}
\end{aligned}
$$

Then

$$
Var\left[\Delta\hat{y}(t_a,t_b)|T\right] = \frac{1}{q_b^2}\sum_{i=1}^{q_b}\lambda_i' V_i \lambda_i. \tag{2.12}
$$

A consistent estimate of the variance is the sample variance of the independent estimated differences, $\Delta\hat{y}_i(t_a, t_b) = \hat{rm}_i(t_b) - \hat{rm}_i(t_a)$. So,

$$\widehat{Var}\left(\Delta\hat{y}(t_a, t_b)|T\right) = \frac{1}{q_b - 1}\sum_{i=1}^{q_b}\left(\Delta\hat{y}_i(t_a, t_b) - \Delta\hat{y}(t_a, t_b)\right)^2 . \tag{2.13}$$

Table 2.3 gives estimates of average labor hours used within 50,000 mile intervals for the complete set of trucks. Note how standard errors increase for higher mileage intervals.

Table 2.3 Estimated average increases in labor hours and standard errors

Mileage interval	Increase in labor hours	Standard error
0-50,000	30.4	.53
50,000-100,0000	29.7	.43
100,000-150,000	33.3	.49
150,000-200,000	36.7	.47
200,000-250,000	39.2	.52
250,000-300,000	41.2	.59
300,000-350,000	42.3	.57
350,000-400,000	44.3	.64
400,000-450,000	46.0	.65
450,000-500,000	48.0	.76
500,000-550,000	47.2	.79
550,000-600,000	47.2	.91
600,000-650,000	49.2	1.11
650,000-700,000	49.4	1.31
700,000-750,000	47.7	1.31
750,000-800,000	47.2	1.55
800,000-850,000	46.8	1.73
850,000-900,000	48.8	2.15
900,000-950,000	47.2	2.46
950,000-1,000,000	45.7	2.70
1,000,000-1,050,000	48.7	4.42
1,050,000-1,100,000	42.1	5.25
1,100,000-1,150,000	33.1	3.95

Selecting the number of pace categories using cross validation

The pace for a truck is the average number of miles per year the truck is driven. As described earlier, this variable can be incorporated into the model by stratifying the trucks into different pace categories and then fitting the model to each pace category. Results for $\Delta\hat{y}$ using four pace categories have been presented earlier. The question now arises, "What is the optimal number of pace categories for estimating increases in cumulative labor hours?"

Let m = the number of pace categories. Let $\Delta\hat{y}_m(t_a, t_b)$ be $\Delta\hat{y}(t_a, t_b)$ with the subscript m to denote the dependence of $\Delta\hat{y}(t_a, t_b)$ on pace category. Let:

$$
\begin{aligned}
n &= \text{number of observations in the data set (5344).} \\
p &= \text{number of trucks in the data set (1182).} \\
n_i &= \text{number of observations for the } i^{th} \text{ truck.} \\
\Delta y_m(t_{i,j}, t_{i,j+1}) - \Delta\hat{y}_m(t_{i,j}, t_{i,j+1}) &= \text{the difference between the observed increase} \\
&\quad \text{and the predicted increase in labor hours} \\
&\quad \text{between two successive observations, } t_{i,j} \text{ and } t_{i,j+1}. \\
&\quad \text{i.e., the residual} \\
\mu(\Delta y_m) &= \text{the mean increase in labor hours} \\
&\quad \text{between } t_{i,j} \text{ and } t_{i,j+1} \\
\bar{s}_{\Delta\hat{y}} &= \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left[\widehat{Var}(\Delta\hat{y}_m(t_{i,j}, t_{i,j+1}))\right]^{\frac{1}{2}} \\
&= \text{the average standard error of } \Delta\hat{y}.
\end{aligned}
$$

Increasing the number of pace categories will reduce the average of the sum of the squared residuals , but the standard error of Δy increases as trucks are subdivided into finer and finer categories.

To illustrate this the model was fitted to six different partitions into pace categories. The partitions are nested and are shown in Table 2.4. Results for the six partitions are given in Table 2.5. As the number of pace categories increases the average of the squared residuals

$$r_m = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=0}^{n_i}(\Delta\hat{y}_m(t_{i,j}, t_{i,j+1}) - (\Delta y_m(t_{i,j}, t_{i,j+1})))^2$$

decreases indicating that predicted increases are closer to the observed increases. However, $\bar{s}_{\Delta\hat{y}}$ increases as m increases, indicating that the standard errors of $\Delta\hat{y}$ are becoming larger.

Table 2.4 Stratification with respect to pace (thousands of miles per year)

$m = 1$	$m = 3$	$m = 6$	$m = 10$	$m = 12$	$m = 24$
No	0-90	0-50	0-50	0-30	0-20
partitioning	90-170	50-90	50-70	30-50	20-30
	> 170	90-130	70-90	50-70	30-40
		130-170	90-110	70-90	40-50
		170-210	110-130	90-110	50-60
		> 210	130-150	110-130	60-70
			150-170	130-150	70-80
			170-190	150-170	80-90
			190-210	170-190	90-100
			> 210	190-210	100-110
				210-230	110-120
				> 230	120-130
					130-140
					140-150
					150-160
					160-170
					170-180
					180-190
					190-200
					200-210
					210-220
					220-230
					230-240
					> 240

Table 2.5 Average of squared residuals and average standard error for the stratifications

	$m=1$	$m=3$	$m=6$	$m=10$	$m=12$	$m=24$
$\sum(\Delta y-\Delta\hat{y})^2/n$	1536.47	1434.23	1381.49	1372.72	1349.34	1318.08
$\bar{s}_{\Delta\hat{y}}$	1.47	2.24	2.89	3.71	3.76	4.97

A reasonable criterion for selecting m is to use the mean squared error of prediction. Conditional on T, the mean squared error between two successive time points is given as

$$\begin{aligned} MSE(\Delta\hat{y}_m(t_{i,j},t_{i,j+1})) &= E\left((\Delta\hat{y}_m(t_{i,j},t_{i,j+1}))-(\mu(\Delta y_m))\right)^2 \\ &= Var(\Delta\hat{y}_m(t_{i,j},t_{i,j+1}))+\left[Bias(\Delta\hat{y}_m(t_{i,j},t_{i,j+1}))\right]^2. \end{aligned} \tag{2.14}$$

The total mean squared error is given by

$$\begin{aligned} C &= E\left[\sum_{i=1}^{p}\sum_{j=0}^{n_i}\left((\Delta\hat{y}_m(t_{i,j},t_{i,j+1}))-(\mu(\Delta y_m))\right)^2\right] \\ &= \sum_{i=1}^{p}\sum_{j=0}^{n_i}\left[Var(\Delta\hat{y}_m(t_{i,j},t_{i,j+1}))+\left[Bias(\Delta\hat{y}_m(t_{i,j},t_{i,j+1}))\right]^2\right]. \end{aligned} \tag{2.15}$$

An approximation to C is

$$d_m=\frac{1}{n}\sum_{i=1}^{p}\sum_{j=0}^{n_i}\left(\Delta\hat{y}_m(t_{i,j},t_{i,j+1})-(\mu(\Delta y_m))\right)^2.$$

A strategy is then to select m so that d_m is minimized. It would seem that d_m could be estimated by the average of the squared residuals, r_m, but the results of Table 2.5 show that r_m decreases as stratification becomes finer. Minimizing r_m would lead to stratification by individual truck.

The problem is that the observation $\Delta y_m(t_{i,j},t_{i,j+1})$ is used in $\Delta\hat{y}_m(t_{i,j},t_{i,j+1})$ to predict itself. What is needed are a set of residuals where $\Delta y_m(t_{i,j},t_{i,j+1})$ and $\Delta\hat{y}_m(t_{i,j},t_{i,j+1})$ are independent. This can be achieved by splitting the data into two groups, a fitting sample and a validation sample. The model is fitted using the data in the fitting sample. The fitted model is then used to estimate increases in the validation sample.

A cross validation method was applied to the truck data by randomly dividing the 1182 trucks into ten (almost) equal sized groups. The first group (the validation sample) was set aside and the model was fitted to the other nine groups. The residuals for the validation sample were then computed using the fitted model. This process was repeated ten times, allowing the second group to be the validation sample and using the other nine groups to fit the model, and so on. Results for the six stratifications are shown in Table 2.6.

Table 2.6 Average squared residual using cross-validation for the six stratifications

	$m=1$	$m=3$	$m=6$	$m=10$	$m=12$	$m=24$
$\sum(\Delta y - \Delta\hat{y})^2/n$	1548.75	1460.65	1438.55	1452.95	1443.64	1474.35

The minimum of d_m for this set of nested stratification schemes occurs at $m = 6$. An interesting result in Table 2.6 is that d_m increases for m=10 and then decreases for $m = 12$. To make sure that this result did not arise from the way the trucks were divided into the ten groups, the procedure was repeated using another random division into ten groups and the same pattern emerged. The stratification scheme for $m = 10$ splits the middle four categories in $m = 6$ but leaves the first category and the last category alone. The stratification scheme for $m = 12$ splits only the first and last categories for $m = 10$. It appears that stratifying very low pace and very high pace trucks has more effect on decreasing d_m than stratifying trucks in the middle range paces. To check this, the lowest pace and the highest pace categories for the $m = 6$ stratification scheme were split further producing an $m = 8$ statification scheme with categories (in thousands of miles per year) 0-30, 30-50, 50-90, 90-130, 130-170, 170-210, 210-230, >230. The value of d_m for this stratification scheme was 1429.24, the lowest of any scheme examined. For this data set, the $m = 8$ stratification scheme seems to be a reasonable partition. It should be noted that the set of nested stratification schemes presented in Table 2.4 is not the only set of schemes possible and so the selection of m will depend on the schemes

actually examined.

Conclusion

The interpolation method is a very simple and easy to compute estimate of $m(t)$. It has a number of practical advantages for estimating a mean growth curve or mean cumulative maintenance cost or labor hours curve where repeated measurements are made on different independent units at different irregular schedules.

- The simplicity of the computations makes the method easy to implement in practice. The estimates can be evaluated using PROC MEANS and a series of data steps in SAS. It is also easy to implement on a spreadsheet program such as LOTUS.

- Estimates for incremental labor hours, $\Delta\hat{y}(t_a, t_b)$, are constrained to be positive for any interval $[t_a, t_b]$.

- Standard errors for $\hat{m}(t)$ and $\Delta\hat{y}(t_a, t_b)$ are easy to evaluate.

- The method is "self-updating". If the analysis is performed frequently it is easy to simply add the new data and run the program to get the new estimates. With more complicated parametric methods there may be the need to search for a new form of the model when new data becomes available. For truck maintenance a new model may be needed to accomodate improvements in truck construction and changes in management. This advantage is also provided by other non-parametric methods to some degree, although some choices such as bandwidth for kernel regression or the number or nearest neighbors for k-NN regression still need to be made.

- The method does not require the complete history of each unit. Recall the estimate of the mean growth curve at t_a used trucks that were driven at least t_a miles and that for these trucks only the time points immediately before and immediately after

t_a were needed (see Equation 2.2). The advantage is that if the firm's management of the maintenance centers has improved through the years, then older data on trucks which could bias results can be discarded.

The linear interpolation method does have disadvantages that need to be considered.

- The method does not provide estimates for the growth curves for each individual unit. Parametric methods such as the mixed model analysis considered in Chapter 4 provide estimates of different growth curves for different individual units.
- The use of covariates to improve predictions, such as pace, requires stratification of the trucks into groups.
- If the curvature of the mean growth curve is large between time points, estimates will be biased.
- The method cannot be used to extrapolate beyond the range of the data. One strategy to obtain predictions beyond the range of the data would be to fit a straight line to the last two observations for a unit and just extrapolate to obtain a prediction. Rao [24] considered such a strategy and found that it performed well compared against more complex methods.

In summary, the linear interpolation method is an attractive procedure for computing nonparametric estimates of an average growth curve when different units are measured on different irregular inspection schedules. It should perform quite well for relatively smooth and slowly changing curves. Other nonparametric methods, such as the locally weighted regression precedure discussed in Chapter 3, may provide more effective ways of incorporating covariates. The mixed model parametric methods described in Chapter 4 provides individual curves for individual units.

3 ESTIMATION AND PREDICTION OF LABOR HOURS USING LOCALLY WEIGHTED REGRESSION

Introduction

Locally weighted regression (loess), introduced by Cleveland [4], is a nonparametric method of fitting a regression surface to data by local fitting of linear or low order polynomial functions of the independent variables. A loess function is contained in the S-Plus package [2]. The loess strategy has not been adapted to data from repeated measures from multiple units where correlated errors are present. In this chapter we apply the loess method for independent observations to repeated measures data to estimate the mean growth curve and then use bootstrap methods to obtain standard errors that account for correlations among repeated measures.

Estimating the regression surface using loess

Loess is one of a number of non-parametric regression methods that can be used to estimate the mean growth curve. Two other commonly used methods are kernel and spline estimation. Diggle, Liang, and Zeger [7] used all three methods to model the mean response in a longitudinal data set and found that all three methods give qualitatively similar results. Müller [18] found that under certain conditions, most importantly equally spaced time points, kernel methods and loess are asymptotically equivalent. In his discussion, Müller compared the two methods for a finite sample and found that there was not much practical difference between the two.

To introduce the loess smoothing procedure assume the model from Chapter 2 with the simplifying assumption of independent errors with constant variance:

$$y_{i,j} = m(t_{i,j}) + \epsilon_{i,j} \qquad \begin{matrix} i = 1, ..., p \\ j = 1, ..., n_i \end{matrix} \tag{3.1}$$

where:

1. m is an unknown smooth regression function
2. $\epsilon_{i,j}$ are observation errors.
3. $E(\epsilon_{i,j}) = 0$
4. $Var(\epsilon_{i,j}) = \sigma^2 < \infty$
5. $Cov(\epsilon_{i,j}, \epsilon_{i,k}) = 0$
6. $N = \sum n_i$, the total number of observations

We will first describe the basic idea behind loess. Let W be a weight function with the following properties:

1. $W(x) \geq 0$ for $|x| < 1$
2. $W(-x) = W(x)$
3. $W(x)$ is a nonincreasing function for $x \geq 1$

Let $0 < \alpha \leq 1$ and let q be αN truncated to an integer. Basically, the procedure goes as follows. For a given t, weights, $w_{i,j}(t)$ are defined for all $t_{i,j}$, $i = 1, \ldots, p$, $j = 1 \ldots, n_i$ using the weight function W. This is achieved by centering W at t and then scaling it so the first point at which W becomes 0 is the q^{th} nearest neighbor of t. The estimated value $\hat{m}(t)$ is the value of a linear or low order polynomial fitted to the data using weighted least squares with the weights $w_{i,j}(t)$.

We now give some of the specific details of the procedure. The smoothness of the loess fit depends on α, the neighborhood parameter, and λ, the degree of the polynomial fitted at t. Let $\Delta_{i,j}(t) = |t - t_{i,j}|$and let $\Delta_{(ij)}(t)$ be the values of these distances ordered from smallest to largest. Let:

$$T(x) = \begin{cases} (1-(x)^3)^3 & 0 \leq x < 1 \\ 0 & x \geq 1 \end{cases}$$

be the tricube weight function. The weight assigned to $(t_{i,j}, y_{i,j})$ in the estimation of the regression curve at some point t is

$$w_{i,j}(t) = T\left(\frac{\Delta_{i,j}(t)}{\Delta_{(q)}(t)}\right).$$

These weights are monotonically decreasing as $t_{i,j}$ increases in distance from t.

Now for each observation, the estimates $\hat{\beta}_k(t), k = 0, 1, ..., \lambda$ for the parameters of a polynomial regression of degree λ are computed using weighted least squares. The $\hat{\beta}_k(t)$ are the values that minimize

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i} w_{i,j}(t)(y_{i,j} - \beta_0 - \beta_1 t - \cdots - \beta_\lambda t^\lambda)^2.$$

Using matrix notation, let:

W_t = the $N \times N$ diagonal matrix of weights, $w_{i,j}(t)$

X = an $N \times (k+1)$ design matrix of the observed times for polynomial regression of degree λ.

y = an $N \times 1$ matrix of the observed responses.

β_t = a $(\lambda+1) \times 1$ vector of regression coefficients for the regression at t.

Then $\hat{\beta}_t = (X'W_tX)^{-1}X'W_ty$, and if we let $x_t' = [1, t, t^2, ..., t^k]$, the loess estimate of the regression curve at t is $\hat{m}(t) = \hat{y}_t = x_t'\hat{\beta}_t$. The variance of this estimate is

$$\begin{aligned} Var(\hat{m}(t)) &= Var(x_t'\hat{\beta}_t) \\ &= \sigma^2 x_t'(X'W_tX)^{-1}X'W_tW_t'X(X'W_tX)^{-1}x_t. \end{aligned} \tag{3.2}$$

Recall from Chapter 2 that we are also interested in estimating the quantity

$$\Delta y(t_a, t_b) = m(t_b) - m(t_a).$$

An estimate of this quantity obtained using local regression is the difference in the estimates of the regression curve at t_a and t_b, i.e.

$$\Delta\hat{y}(t_a, t_b) = \hat{m}(t_b) - \hat{m}(t_a). \tag{3.3}$$

Calculating the variance of $\Delta\hat{y}(t_a, t_b)$ is more complicated because $\{w_{i,j}(t_b)\}$, the weights used for the local regression at t_b, are not equal to $\{w_{i,j}(t_a)\}$,the weights used at t_a. Hence, $\hat{\beta}_{t_b}$ is generally not equal to $\hat{\beta}_{t_a}$. Thus,

$$\begin{aligned} Var(\Delta\hat{y}(t_a, t_b)) &= Var(\hat{m}(t_b) - \hat{m}(t_a)) \\ &= Var(x_{t_a}'\hat{\beta}_{t_a}) + Var(x_{t_b}'\hat{\beta}_{t_b}) - 2Cov(x_{t_b}'\hat{\beta}_{t_b}, x_{t_a}'\hat{\beta}_{t_a}) \\ &= \sigma^2 x_{t_a}'(X'W_{t_a}X)^{-1}X'W_{t_a}W_{t_a}'X(X'W_{t_a}X)^{-1}x_{t_a} \\ &\quad +\sigma^2 x_{t_b}'(X'W_{t_b}X)^{-1}X'W_{t_b}W_{t_b}'X(X'W_{t_b}X)^{-1}x_{t_b} \\ &\quad -2\sigma^2 x_{t_b}'(X'W_{t_b}X)^{-1}X'W_{t_b}W_{t_a}'X(X'W_{t_a}X)^{-1}x_{t_a} \\ &= \sigma^2\left[x_{t_b}'(X'W_{t_b}X)^{-1}X'W_{t_b} - x_{t_a}'(X'W_{t_a}X)^{-1}X'W_{t_a}\right] \\ &\quad \times\left[W_{t_b}'X(X'W_{t_b}X)^{-1}x_{t_b} - W_{t_a}'X(X'W_{t_a}X)^{-1}x_{t_a}\right]. \end{aligned} \tag{3.4}$$

Nonconstant variance

The loess fitting procedure can be modified to accommodate nonconstant variance. Following Cleveland [6] assume there is a known set of constants $\{a_{i,j}\}$ such that

$Var(\sqrt{a_{i,j}}\epsilon_{i,j}) = \sigma^2 < \infty$. Then the weight for $(t_{i,j}, y_{i,j})$ becomes $a_{i,j}w_{i,j}(t)$. Let $A =$ denote an $N \times N$ diagonal matrix of weights $\{a_{i,j}\}$ and let $B_t = W_t A$. Then, we have

$$\hat{\beta}_t = (X'B_tX)^{-1}X'B_ty \tag{3.5}$$

and the results in the previous section hold with W_t replaced by B_t.

Correlated errors

The development of the loess procedure up to this point has assumed $Cov(\epsilon_{i,j}, \epsilon_{i,k}) = 0$. Now suppose $Var(y) = \sigma^2 V$ where V is a known positive definite matrix. Consider the transformations $z = V^{-\frac{1}{2}}y$ and $U = V^{-\frac{1}{2}}X$. Then $Var(z) = \sigma^2 I$. Applying the weight matrix W_t to the transformed data gives

$$\hat{\beta}_t = (X'V^{-\frac{1}{2}}W_tV^{-\frac{1}{2}}X)^{-1}X'V^{-\frac{1}{2}}W_tV^{-\frac{1}{2}}y. \tag{3.6}$$

Equations 3.2 and 3.4 still hold with W_t replaced by $V^{-\frac{1}{2}}W_tV^{-\frac{1}{2}}$.

In practice, the covariance matrix V is not known. Estimating V from the truck data would be difficult because there are no replications at the various time points, t. It should also be noted that the present implementation of loess in S-Plus does not accomodate the assumption of $Var(y) = \sigma^2 V$ unless V is a diagonal matrix. Our strategy will be to use Equation 3.5 to calculate $\hat{\beta}_t$. The loss in efficiency should be small since only observations within a truck are correlated. The computation of $\hat{\beta}_t$ uses at most two or three time points from a given truck since only a portion of the data, depending on α, is used in the computation. Thus, for a given t, any observation will be correlated with at most one or two other observations used in the calculation of $\hat{\beta}_t$.

Application to the truck data

We now apply the loess smoothing procedure to the truck data set. Recall $y_{i,j} =$ cumulative labor hours and $t_{i,j} =$ cumulative mileage for the i^{th} truck at the end of

the j^{th} year of service. Furthermore, Figure 1.1 shows that the variance homogeneity assumption is violated, so we will use 3.5. We will ignore potential correlation among repeated observations on a single truck in fitting the curve, but we will later use bootstrap methods to compute standard errors that account for such correlations. The basic steps in applying the procedure are:

1. Estimate the weights, $a_{i,j}$.

2. Select a value for λ, the degree of polynomial fitted locally.

3. Select a value for α, the neighborhood size. α is sometimes called the span.

Determination of weights, $\{a_{i,j}\}$

Estimation of the weights $\{a_{i,j}\}$ would present no problem if there were multiple observed y's at each observed value of t, as is the case in growth curve studies with regular inspections. In that case, one would use the sample variance at each value of t. The weights would be inversely proportional to the sample variances. However, the truck maintenance data generally has only one observed cumulative labor hour value for each cumulative mileage value. To estimate the weights for the truck maintenance data we proceed as follows. The range of cumulative mileage values, 0 to 1,200,000 miles, was divided into 12 classes of equal length, 0-100,000 miles, 100,000-200,000 miles and so forth. The sample standard deviation for cumulative labor hours and the average cumulative mileage was calculated for each class These are presented in Table 3.1.

Figure 3.1 is a plot of standard deviation of cumulative labor hours versus average cumulative mileage. A strong linear relationship is apparent with the possible exception of class 12, which contains relatively few trucks.

The regression of standard deviation of cumulative labor hours on average cumulative mileage, weighted by the number of observations in each class, gave the estimated

Table 3.1 Standard deviation of cumulative labor hours by class

Class	Cumulative Mileage	Standard Deviation	Average Cumulative Mileage	Number of Units
1	0-100,000	33.62	53510.13	789
2	100,000-200,000	50.74	148045.44	926
3	200,000-300,000	69.96	248539.32	792
4	300,000-400,000	82.35	349395.17	750
5	400,000-500,000	105.97	448427.47	675
6	500,000-600,000	112.76	547960.05	565
7	600,000-700,000	135.91	643476.06	375
8	700,000-800,000	152.01	745849.17	208
9	800,000-900,000	158.06	844967.51	112
10	900,000-1,000,000	182.79	950765.79	75
11	1,000,000-1,100,000	221.91	1048932.40	58
12	1,100,000-1,200,000	170.07	1133209.32	19

regression line

$$\text{standard deviation} = 26.250 + .000167(\text{average cumulative mileage})$$

with r^2=.986. A reasonable estimate of $\sigma(t_{i,j})$ would then be:

$$\hat{\sigma}(t_{i,j}) = 26.250 + .000167(t_{i,j}) \tag{3.7}$$

and the weights are then estimated by:

$$\hat{a}_{i,j} = \frac{1}{\hat{\sigma}^2(t_{i,j})}. \tag{3.8}$$

Selecting a value for λ, the degree of polynomial

Cleveland [4] recommends choosing $\lambda = 1$ as striking a good balance between computational ease and the need for flexibility to reproduce patterns in the data. When the regression surface has substantial curvature, such as a local maximum or minimum, the choice $\lambda = 2$ may be more appropriate. For the truck data, the scatterplot of the observations shown in Figure 1.1 and results of Chapter 2 suggest that $m(t)$ does not have a local maximum or minimum, giving more support for the choice $\lambda = 1$.

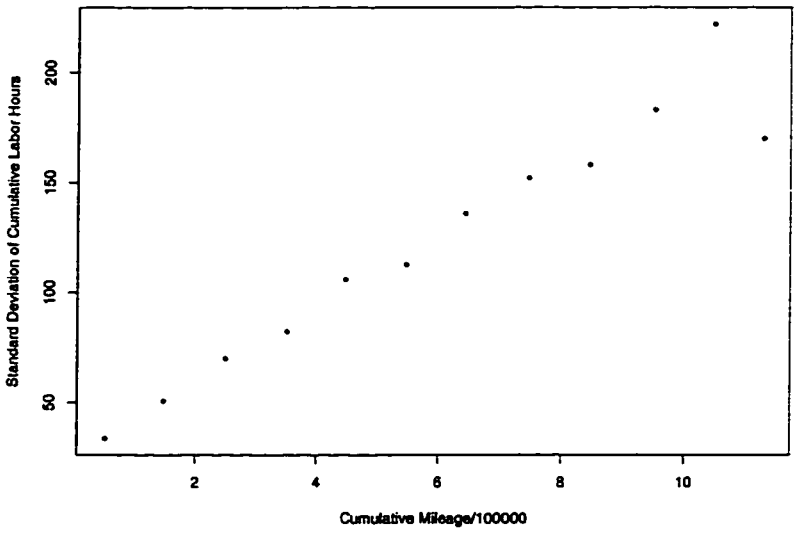

Figure 3.1 Standard deviation of cumulative hours versus cumulative mileage

Selecting a value for α using the M-plot

The choice of $0 < \alpha \leq 1$ governs the amount of smoothing by the loess fitting procedure and is similar to the choice of bandwidth in kernel estimation. As the value of α increases the amount of smoothing increases. The usual goal of selecting α is to pick as large a value as possible to minimize the variability of the smoothed points without unduly increasing the bias. In cases where the goal is to use the regression curve for prediction, a smaller value of α which reduces the bias may be more appropriate.

The M-plot, introduced by Cleveland and Devlin [5], is a graphical tool to aid in selecting α. It is an extension of a procedure invented by Mallows called C_p for choosing a subset of the independent variables based on estimates of the mean squared error for each subset. The development here is an extension to the case of nonconstant variance. As before, we assume that there is a known set of constants such that $Var(\sqrt{a_{i,j}}\epsilon_{i,j}) = \sigma^2 < \infty$. We are also assuming that $Cov(\epsilon_{i,j}, \epsilon_{i,k}) = 0$, even though observations within a truck are likely to be correlated.

The loess estimate of the regression curve at t is

$$\hat{m}(t) = x_t'(X'B_tX)^{-1}X'W_tAy \tag{3.9}$$

where

$$\begin{aligned} A &= \text{the } N \times N \text{ diagonal matrix of weights } \{a_{i,j}\}. \\ W_t &= \text{the } N \times N \text{ diagonal matrix of weights } \{w_{i,j}(t)\}. \\ B_t &= W_tA. \\ x_t' &= [1, t, t^2, \cdots, t^k]. \end{aligned}$$

Let $l(t) = x_t'(X'B_tX)^{-1}W_t$ be a $1 \times N$ matrix. Then $\hat{m}(t) = l(t)Ay$. Now let $\hat{y}_{i,j} = \hat{m}(t_{i,j})$ be the fitted values, $\hat{\epsilon}_{i,j} = y_{i,j} - \hat{y}_{i,j}$ be the residuals, y be the vector of observed cumulative labor hours, $\hat{y}$ be the vector of fitted values and $\hat{\epsilon}$ be the vector of residuals. Then $\hat{y} = LAy$ and $\hat{\epsilon} = (I - LA)y$ where L is an $N \times N$ matrix with rows $l(t)$ and I is the $N \times N$ identity matrix. It follows that

$$Var(\hat{y}) = Var(LAy) = \sigma^2 LA(A^{-1})AL' = \sigma^2 LAL' \tag{3.10}$$

and

$$Var(\hat{\epsilon}) = Var((I - LA)y) = \sigma^2(I - LA)A^{-1}(I - LA)'. \tag{3.11}$$

If $\hat{m}(t)$ is an unbiased estimator of $m(t)$, then

$$E\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} a_{i,j}\hat{\epsilon}_{i,j}^2\right) = E\left(\hat{\epsilon}'A\hat{\epsilon}\right) = \sigma^2 tr\left[(I - AL')(I - AL)\right] \tag{3.12}$$

and an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'A\hat{\epsilon}}{tr\left[(I - AL')(I - AL)\right]}. \tag{3.13}$$

The expected mean squared error summed over the observed cumulative mileages and divided by σ^2 is

$$M_\alpha = \frac{E\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\hat{m}_\alpha(t_{i,j}) - m(t_{i,j}))^2}{\sigma^2}$$

where the subscript α denotes dependence on the choice of span. M_α consists of a bias term and a variance term, i.e.,

$$M_\alpha = B_\alpha + V_\alpha$$

where

$$B_\alpha = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(E(\hat{m}_\alpha(t_{i,j})) - m(t_{i,j}))^2}{\sigma^2}$$

$$V_\alpha = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i} Var(\hat{m}_\alpha(t_{i,j}))}{\sigma^2} = trL_\alpha AL'_\alpha.$$

Let RSS_α be the residual sum of squares of $\hat{m}_\alpha(t)$ using the rescaled residuals $\sqrt{a_{i,j}}\epsilon_{i,j}$. Then

$$\frac{E(RSS_\alpha)}{\sigma^2} = \frac{E(\hat{\epsilon}'A\hat{\epsilon})}{\sigma^2} = [tr(I - AL'_\alpha)(I - AL_\alpha)] + B_\alpha.$$

If σ^2 were known an estimate of M_α would be

$$M_\alpha = \frac{\hat{\epsilon}'A\hat{\epsilon}}{\sigma^2} - tr(I - AL'_\alpha)(I - AL_\alpha) + trL_\alpha AL'_\alpha.$$

Now let $\hat{\sigma}_s^2$ be the estimate for σ^2 where s is a small span. If the span is small enough, then $\hat{m}_s(t)$ is nearly unbiased resulting in a nearly unbiased estimate of σ^2 using Equation 3.13. The estimate of M_α becomes

$$\hat{M}_\alpha = \frac{\hat{\epsilon}'A\hat{\epsilon}}{\hat{\sigma}_s^2} - tr(I - AL'_\alpha)(I - AL_\alpha) + trL_\alpha AL'_\alpha.$$

The M-plot is a plot of $\hat{M}_\alpha$ versus V_α for a selection of various values of α between s and 1. Usually, as α increases, the bias increases and the variance decreases. Using the M-plot allows us to see the trade-off between bias and variance.

The development of $\hat{M}_\alpha$ did not account for correlated observations. Note that as the neighborhood parameter α decreases the number of observations within a truck decreases and so correlation becomes less of a problem.

Figure 3.2 is an M-plot with s taken to be 0.05. The rightmost point is $\alpha = 0.05$ and the line is $\hat{M}_\alpha = V_\alpha$. Vertical distances from the line reflect the size of the bias term in

$\hat{M}_\alpha$. The rightmost point lies on the line since we assumed that the bias for $\alpha = 0.05$ is negligible. The criterion used by Cleveland and Devlin [5] and Cleveland, Devlin and Grosse [6] in selecting α was to choose α approximately where M_α begins a dramatic rise. In this way variance is minimized without introducing undue bias. After examining the M-plot, α was chosen to be 0.35 since there is a rapid rise in the bias for α values greater than 0.35.

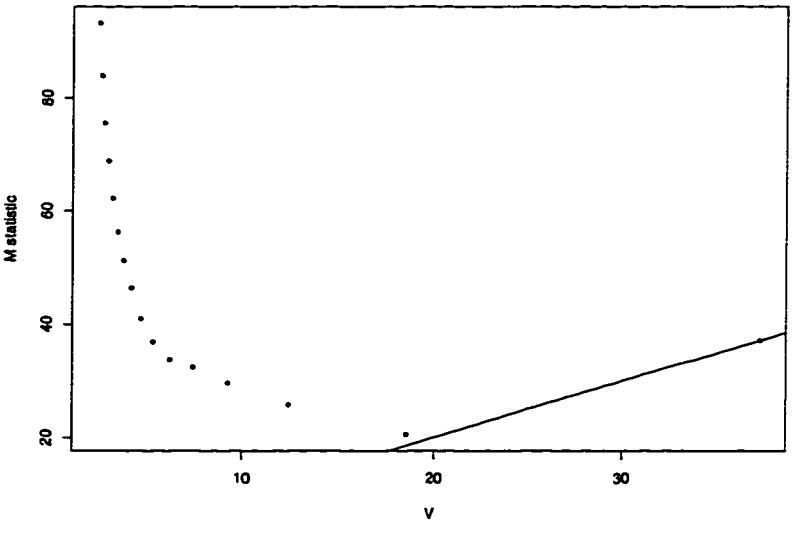

Figure 3.2 M-plot of M statistic versus V for $\alpha = 0.05$ to $\alpha = 0.80$ in steps of 0.05

Computation of $\hat{m}(t)$

Using the loess function in S-Plus, a loess model with the specifications $\alpha = 0.35$, $\lambda = 1$, and weights $\hat{a}_{i,j}$ was fit to the truck data. A plot of cumulative labor hours versus $\hat{m}(t)$ is shown in Figure 3.3. Loess models with α values between 0.25 and 0.50 produced plots very similar to Figure 3.3. As α decreases below 0.25, the plots become more irregular and are not monotonically increasing. A plot of $\hat{m}(t)$ from the interpolation approach used in the previous chapter is overlaid on $\hat{m}(t)$ computed from loess in Figure 3.4

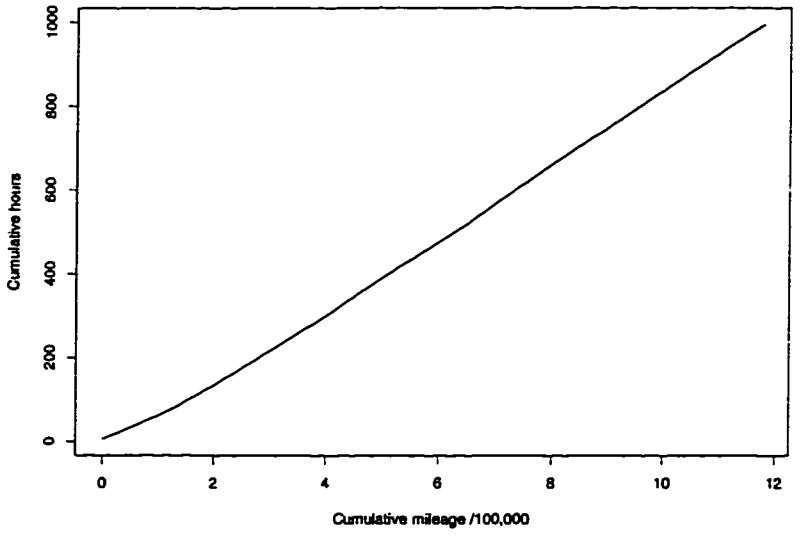

Figure 3.3 Estimated cumulative labor hours vs. cumulative mileage

Diagnostics

We now examine some diagnostic plots to determine if the choices for $\alpha = 0.35$, $\lambda = 1$, and weights $\hat{a}_{i,j}$ are reasonable. Let $\hat{\epsilon}_{i,j} = y_{i,j} - \hat{m}(t_{i,j})$. To investigate distributional assumptions we use the rescaled residuals:

$$\hat{\epsilon}^*_{i,j} = \sqrt{a_{i,j}}\hat{\epsilon}_{i,j}.$$

A plot of the standardized residuals versus cumulative labor hours is shown in Figure 3.5. A loess smoothed curve is added to the residual plot to aid in the detection of any patterns. No effect appears to be present, so the estimated loess curve appears to adequately model the information on cumulative labor hours provided by cumulative mileage. The overall pattern of the residuals still shows a problem with nonconstant variance, but it is an improvement over the loess model with no weights. See Figure 3.6 for a residual plot from a loess model without the weights, $\hat{\alpha}_{i,j}$.

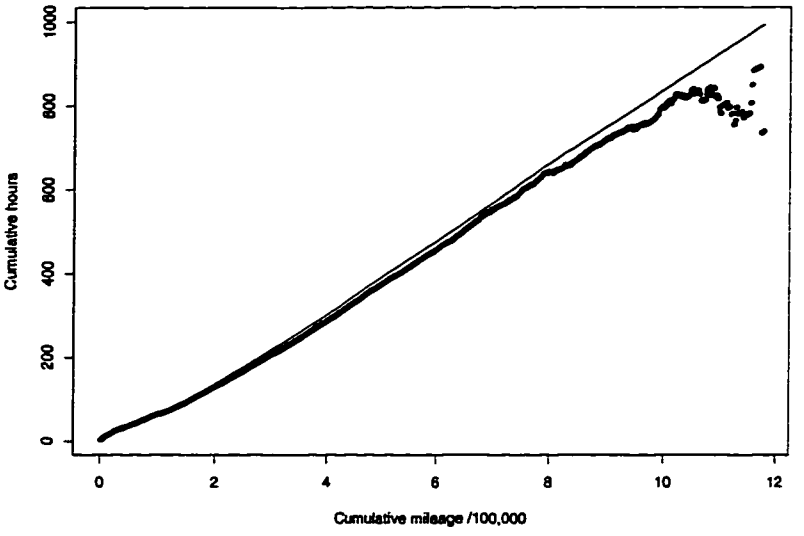

Figure 3.4 Estimated cumulative labor hours vs. cumulative mileage – loess and interpolation approach

Bootstrap estimates for standard error of $\hat{m}(t)$ and $\Delta\hat{y}(t_a, t_b)$

The expressions given for $Var(\hat{m}(t))$ and $Var(\Delta\hat{y}(t_a, t_b))$ given by Equations (3.2) and (3.4) do not account for the correlation of errors within trucks. The derivation of more appropriate variance formulas would require specification of a model for within truck correlations as a function of a few additional parameters that would be estimated from correlation among residuals. Alternatively, a bootstrap procedure can be used to estimate variances and covariances of estimates of $\hat{m}(t)$ and $\Delta\hat{y}(t_a, t_b)$. The bootstrap algorithm for calculating standard errors was carried out as follows, using the current implementation of the loess function is S-Plus:

1. There are a total of 5344 observations from 1182 trucks. To account for correlation of errors within a truck we used the entire set of observations within a truck (truck history) rather than individual observations for the sampling units. A total of 1000 independent bootstrap samples each consisting of 1182 resampled truck histories was taken.

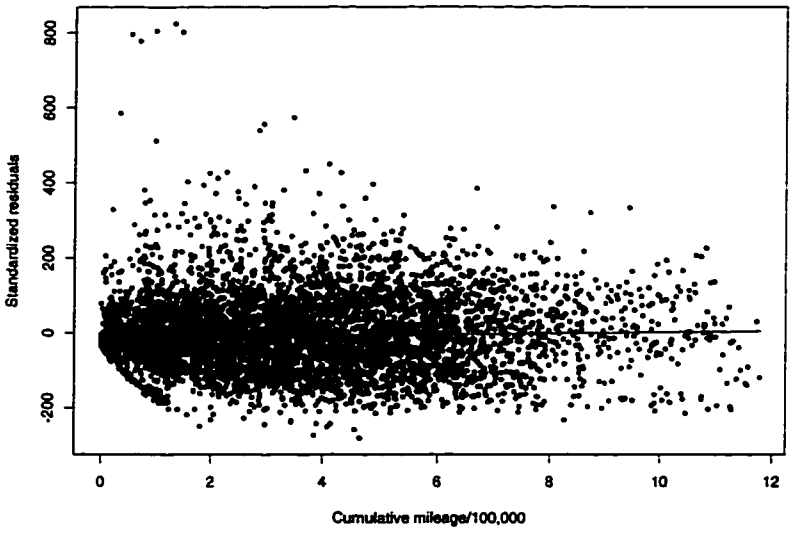

Figure 3.5 Standardized residuals versus cumulative mileage - alpha=0.35

2. Let $\hat{m}^{*i}(t)$ be the value of $\hat{m}(t)$ calculated from the i^{th} bootstrap sample. For each bootstrap replication, the values of $\hat{m}^{*i}(t)$ were calculated at 50,000 mile intervals over the range of t using the S-Plus function *loess*. The range of t was 0 to 1,200,000 miles, so each bootstrap replication gave 25 values of $\hat{m}^{*i}(t)$, i.e. $\hat{m}^{*i}(0)$, $\hat{m}^{*i}(50,000)$, $\hat{m}^{*i}(100,000)$, ... , $\hat{m}^{*i}(1,200,000)$. Let $\hat{M}^{*i}$ be a 25×1 column vector of the values of $\hat{m}^{*i}(t)$ and $\hat{Var}(\hat{M}^{*})$ be a 25×25 variance matrix calculated from the 1000 bootstrap replications $\hat{M}^{*i}$.

Let $\hat{M}$ be a 25×1 column vector of the values of $\hat{m}$ at t = 0, 50,000, 100,000 ,..., 1,200,000. Then $\hat{Var}(\hat{M}^{*})$ is an estimate of $Var(\hat{M})$. The square root of the diagonal elements of $\hat{Var}(\hat{M}^{*})$ are the estimates for the standard deviations of $\hat{m}(t)$ values at 50,000 mile intervals. Computations for the standard errors of$\Delta\hat{y}(t_a, t_b)$ at 50,000 mile intervals are straightforward and the results are shown in Table 3.2 . To get the standard error for $\Delta\hat{y}(t_a, t_b)$ for *any* arbitrary $t_b > t_a$ there are two options:

1. Rerun the bootstrap routine with the particular $[t_a, t_b]$ being considered.

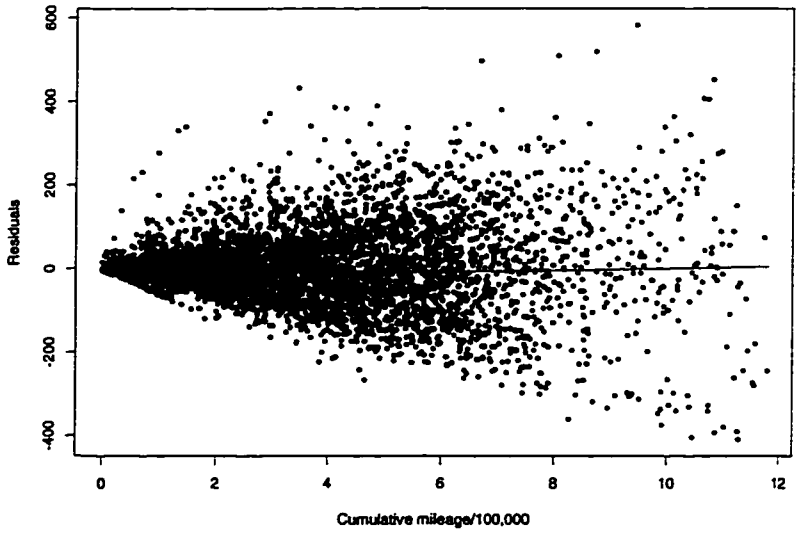

Figure 3.6 Residuals versus cumulative mileage - alpha=0.35, No weights

2. Estimate the variance of $\Delta\hat{y}(t_a, t_b)$ with an estimate of the variance of an approximation to $\Delta\hat{y}(t_a, t_b)$. This estimate is obtained from the 25×25 covariance matrix computed for 50,000 mile intervals using the bootstrap routine. First approximate $\hat{m}(t_a)$ by a linear interpolation between the endpoints of the appropriate 50,000 mile interval. Do the same to approximate $\hat{m}(t_b)$. The resulting approximation to $\Delta\hat{y}(t_a, t_b)$ is a linear combination of the elements of M and may be represented as

$$\tilde{\Delta}\hat{y}(t_a, t_b) = l'\hat{M}$$

Then $Var(l'\hat{M}) = l'Var(\hat{M})l$, which may be estimated by $l'\widehat{Var}(\hat{M}^*)l$. Use this as the variance estimate for $\Delta\hat{y}(t_a, t_b)$.

A comparison of Table 3.2 and Table 2.3 shows that estimated increases for 50,000 mile intervals tend to be larger for the linear interpolation method than for loess. This can be explained by the differences in which the two methods compute estimated increases. Consider the mileage interval (t_a, t_b). The loess estimate of cumulative labor hours at t_b includes all trucks in a neighborhood determined by the span, α. In particular, the loess estimate at t_b will include some trucks with cumulative mileages less

Table 3.2 Estimated increase in labor hours

Mileage interval	Estimated increase in labor hours	Standard error
0-50,000	27.8	.82
50,000-100,000	29.0	.77
100,000-150,000	34.1	.76
150,000-200,000	38.5	.88
200,000-250,000	39.9	1.05
250,000-300,000	41.5	.99
300,000-350,000	42.1	1.12
350,000-400,000	41.9	1.31
400,000-450,000	44.8	1.33
450,000-500,000	44.0	1.41
500,000-550,000	43.6	1.61
550,000-600,000	43.5	1.92
600,000-650,000	44.2	2.08
650,000-700,000	47.2	2.15
700,000-750,000	47.4	2.31
750,000-800,000	46.7	2.51
800,000-850,000	44.7	2.66
850,000-900,000	43.5	2.62
900,000-950,000	43.7	2.53
950,000-1,000,000	44.0	2.50
1,000,000-1,050,000	44.1	2.54
1,050,000-1,100,000	44.0	2.62
1,100,000-1,150,000	43.8	2.70
1,150,000-1,200,000	43.4	2.78

than t_b. Many of these are low pace trucks which have a higher rate of labor hours per mile. The linear interpolation method for estimating increases will exclude these trucks since the method excludes trucks with cumulative mileages less than t_b. The result of including these trucks in the loess computation is that the estimate of cumulative labor hours at t_b will be decreased. The effect becomes more pronounced as t_b increases (by 50,000 mile intervals, for example) since smaller numbers of trucks make it to the higher cumulative mileages. The end result is that estimated increases over 50,000 mile intervals tend to be smaller for loess than for the linear interpolation method.

The main feature of the truck data which causes this problem in estimating increases using loess is that the shape of an individual truck's growth curve is related to the ending of the curve. Low pace trucks tend to have short curves that increase rapidly. High pace trucks will tend to have long curves that increase less rapidly. In short, there is an effect due to pace which contributes to the problem. In the next section we incorporate pace into the loess analysis.

The use of loess to estimate the mean growth curve for a population when time is the only explanatory variable should be handled with some care. In general, the time span covered by an individual's growth curve should not be related to the shape of the individual's growth curve. Examples where this might be the case are data sets from the biological sciences, such as the often cited children's dental study in Potthoff and Roy [19].

A further comparison of Table 3.2 and Table 2.3 shows that standard errors for estimated increases in labor hours tends to be smaller for the linear interpolation method than for loess except for mileages above 950,000 where there are few trucks. As noted before, the standard errors produced by loess are dependent on the choice of α, the neighborhood parameter. A larger α would produce estimates with smaller standard errors but with larger bias.

Incorporating pace into the loess analysis

The variable pace, the average number of miles a truck is driven per year, was introduced in Chapter 2 and was found to improve estimates of cumulative labor hours. In this section we incorporate both cumulative mileage and pace into the loess analysis. Let r_i be the average mileage per year for the i^{th} truck. Then the model given in (3.1) is generalized as

$$y_{i,j} = m(t_{i,j}, r_i) + \epsilon_{i,j}. \tag{3.14}$$

The loess procedure for fitting a smooth surface in two dimensions is a straightforward extension of the procedure for one dimension. As before, the smoothness of the loess fit depends on λ, the degree of polynomial fitted at the point (t, r), and α, the span or smoothness parameter.

We choose $\lambda = 1$ for the same reason as in the one dimensional case, i.e. Cleveland's [4] recommendation to $\lambda = 1$ as striking a good balance between computational ease and the need for flexibility to reproduce patterns in the data. With $\lambda = 1$, the estimate $\hat{m}(t, r)$ at a point (t, r) is a weighted local regression of the form

$$y = \beta_0 + \beta_1 t + \beta_2 r + \beta_3 tr$$

The $\hat{\beta}(t)$ are the values that minimize

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} w_{i,j}(t)(y_{i,j} - \beta_0 - \beta_1 t - \beta_2 r - \beta_3 tr)^2$$

The weights $w_{i,j}(t)$ are determined using the tricube weight function. Distances between points in two dimensions are calculated using a Euclidean distance between rescaled measurements, i.e. the predictors are divided by their standard deviation.

Determination of weights, $\{a_{i,j}\}$

As before, there is still the problem of nonconstant variance. To determine the effect of pace on the variance of y, the trucks were stratified into four pace categories according to Table 2.1. In each pace category, the range of cumulative mileage values, 0 to 1,200,000, was divided into 12 classes of equal length, 0-100,000, 100,000-200,000, and so forth. Thus, the trucks were partitioned into $12 \times 4 = 48$ groups. The sample standard deviation and the average cumulative mileage were computed for each group. Then two regressions were performed, both weighted by the number of observations in each group. The results were:

standard deviation = 23.48 + 0.000163(average cumulative mileage)

$$r^2 = .930$$

$$\text{standard deviation} = 37.89 + 0.000172(\text{average cumulative mileage}) - .000147(\text{pace})$$

$$r^2 = .945$$

Pace does not seem to have a large impact on the prediction of standard deviation. Therefore, the first equation was used to estimate the weights according to

$$\hat{a}_{i,j} = \frac{1}{\hat{\sigma}^2(t_{i,j})}.$$

Selection a value of α using the M-plot

The M-plot, described earlier, was used to select the value of α. In Figure 3.7 the rightmost point is for α=0.05. Beyond α=0.25 there is a rapid rise in the amount of bias. Based on the M-plot, α=0.25 was selected as the span.

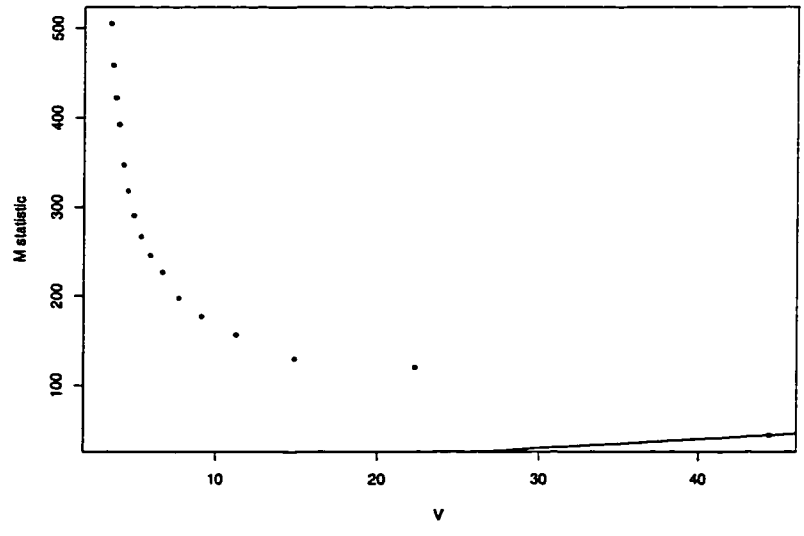

Figure 3.7 M-plot of M statistic versus V for $\alpha = 0.05$ to $\alpha = 0.80$ in steps of 0.05

Computation of $\hat{m}(t,r)$ and bootstrap estimates for standard errors of $\hat{m}(t,r)$ and $\Delta\hat{y}(r,t_a,t_b)$

A loess model with the specifications $\alpha = 0.25$, $\lambda = 1$, and recomputed weights $\hat{a}_{i,j}$ was fit to the truck data. A perspective plot of the surface is shown in Figure 3.8. The grid was constructed for 0 to 1,200,000 cumulative miles with 50,000 increments and for pace values between 25,000 to 300,000 average miles per year with 25,000 miles per year increments. Note that cumulative labor hours have nearly a straight line relationship with cumulative mileage in the higher pace categories but the relationship is more curved and rises more quickly for low pace categories. This relationship is clearly shown in Figure 3.9 where predicted cumulative labor hours versus cumulative mileage is shown for the pace values 50,000, 100,000, 200,000, and 300,000 average miles per year.

The bootstrap algorithm was carried out as previously described except $\hat{m}^{*i}(t,r)$ was calculated on a 12×25=300 point grid for 1000 independent bootstrap samples. Instead of generating a single 25×25 variance matrix as before, we generated twelve 25×25 variance matrices, one for each value of pace in the grid. Estimates of mean labor hours used within 50,000 mile intervals and their standard errors are shown in Tables 3.3, 3.4, and 3.5. Note that within a given mileage interval the standard errors are highest when pace is 25000 miles per year and then decrease as pace increases to about 125,000 to 175,000 miles per year. Standard errors tend to be fairly constant for higher pace values. Standard errors from Table 3.2, where pace was not incorporated into the loess analysis, appear to be "averages" of the standard errors within the mileage intervals from Tables 3.3, 3.4, and 3.5. It should be noted that some of the values in these tables are extrapolations from the observed data. In the pace category 25000 average miles per year, for example, there were no trucks that had cumulative mileage greater than 400,000 miles.

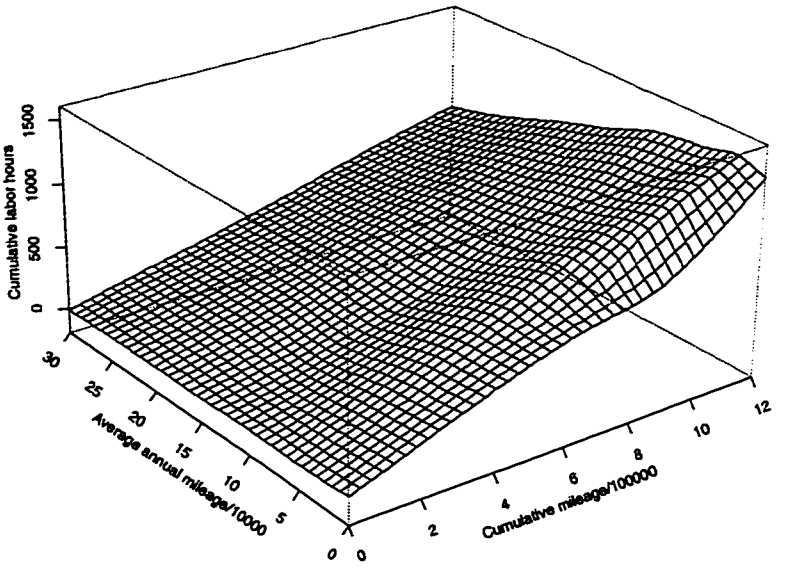

Figure 3.8 Cumulative labor hours versus cumulative mileage and pace

Conclusion

This chapter has considered loess as a nonparametric method for estimating $m(t)$, the overall mean growth curve. Some comparisons with the linear interpolation approach considered in the previous chapter can be made.

- Implementing loess is more complicated since choices about the degree of polynomial to be locally fitted and the neighborhood size need to be made.
- As for linear interpolation, this implementation of loess does not provide estimates of growth curves for individual trucks.
- Compared with the linear interpolation method, standard errors of $\hat{m}(t)$ and $\Delta\hat{y}(t_a, t_b)$ for the loess method tend to be higher for low and moderate values of t where many trucks are available to the linear interpolation method and lower for high values of t where relatively few trucks are available to the linear interpolation method.

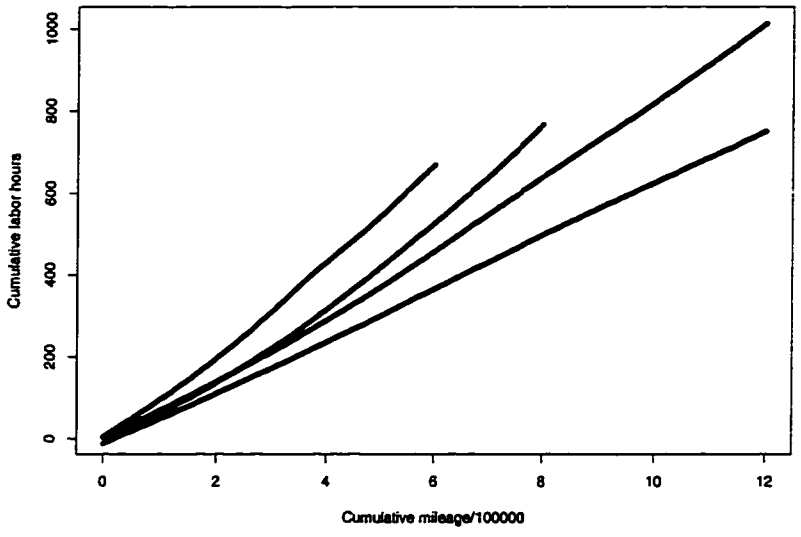

Figure 3.9 Loess curves for pace values 50,000, 100,000, 200,000 and 300,000

- Computation of standard errors for $\Delta\hat{y}(t_a, t_b)$ is more difficult than for the linear interpolation approach. Bootstrap methods were used rather that direct calculation.

- Incorporating covariates such as pace to improve estimates does not require stratification of the trucks into pace categories.

In this chapter we have seen that loess is another alternative for estimating $m(t)$ and for providing estimates and standard errors of $\Delta\hat{y}(t_a, t_b)$. Although loess can be used in this manner, it is also a very useful tool for determining the form of the overall mean growth curve in an exploratory data analysis. This can be helpful in choosing the form of a model with more complicated parametric methods such as the mixed model analysis considered in the next chapter.

Table 3.3 Estimated mean labor hours used within consecutive 50,000 mile intervals. Standard errors are shown in parentheses.

	Average miles per year			
Mileage interval (×1000)	25000	50000	75000	100000
0-50	47.64 (3.16)	44.73 (2.09)	36.30 (1.08)	28.38 (.75)
50-100	48.81 (3.01)	46.77 (2.02)	39.75 (1.15)	30.97 (.69)
100-150	50.39 (3.09)	48.95 (2.09)	44.02 (1.28)	34.19 (.66)
150-200	52.06 (3.39)	51.74 (2.44)	48.26 (1.50)	38.75 (.67)
200-250	53.27 (3.82)	55.31 (3.22)	50.28 (1.72)	41.21 (.74)
250-300	54.43 (4.49)	59.69 (4.47)	51.55 (1.97)	43.14 (.85)
300-350	55.38 (5.59)	61.99 (6.35)	53.95 (2.12)	45.65 (1.12)
350-400	55.18 (7.11)	58.39 (7.45)	55.95 (2.42)	48.75 (1.19)
400-450	54.04 (9.20)	54.86 (6.63)	57.25 (2.69)	51.43 (1.10)
450-500	51.94 (12.65)	57.45 (6.06)	58.24 (3.02)	52.19 (1.47)
500-550	48.29 (15.69)	61.31 (6.13)	60.90 (3.69)	53.28 (1.90)
550-600	47.83 (15.89)	64.01 (6.40)	59.21 (4.05)	54.12 (2.34)
600-650	51.89 (15.70)	65.43 (7.05)	57.36 (4.46)	55.95 (2.92)
650-700	63.01 (14.52)	66.08 (7.77)	54.70 (4.81)	58.57 (3.50)
700-750	68.80 (16.12)	64.37 (8.01)	52.02 (5.14)	61.54 (4.03)
750-800	78.09 (18.57)	60.73 (8.11)	53.67 (5.60)	64.54 (4.23)
800-850	85.10 (19.88)	56.98 (8.13)	55.74 (6.12)	66.96 (4.61)
850-900	85.64 (19.40)	53.76 (7.95)	58.77 (6.38)	69.06 (4.61)
900-950	78.92 (17.68)	51.90 (8.50)	61.78 (6.81)	69.85 (4.36)
950-1,000	70.84 (15.47)	51.92 (8.73)	63.53 (7.26)	70.25 (4.52)
1000-1050	60.67 (13.46)	54.32 (8.86)	67.12 (7.12)	70.63 (4.67)
1050-1100	55.31 (12.02)	57.62 (9.04)	68.73 (6.87)	70.06 (4.65)
1100-1150	52.62 (11.27)	59.49 (9.51)	70.50 (6.79)	68.09 (4.47)
1150-1200	58.82 (11.44)	62.62 (9.50)	71.58 (6.77)	66.90 (4.64)

Table 3.4 Estimated mean labor hours used within consecutive 50,000 mile intervals. Standard errors are shown in parentheses.

	Average miles per year			
Mileage interval (×1000)	125000	150000	175000	200000
0-50	25.27 (.69)	26.26 (.70)	30.74 (.85)	33.11 (.84)
50-100	27.15 (.65)	27.55 (.66)	31.98 (.84)	33.86 (.84)
100-150	29.16 (.67)	29.44 (.66)	33.31 (.85)	34.61 (.84)
150-200	32.40 (.68)	31.95 (.71)	34.68 (.88)	53.43 (.85)
200-250	34.79 (.67)	34.48 (.80)	36.15 (.90)	36.15 (.86)
250-300	36.74 (.80)	36.37 (.92)	37.80 (.94)	36.96 (.88)
300-350	38.35 (.80)	38.78 (.98)	39.46 (1.01)	37.90 (.92)
350-400	38.96 (.84)	40.58 (.99)	40.63 (1.07)	83.94 (.95)
400-450	40.75 (.95)	42.44 (1.02)	41.62 (1.12)	41.01 (1.05)
450-500	43.59 (1.24)	45.04 (1.21)	42.91 (1.05)	41.29 (1.11)
500-550	48.92 (1.67)	47.81 (1.25)	44.64 (.98)	42.53 (1.21)
550-600	55.48 (2.10)	51.56 (1.42)	46.44 (1.02)	44.27 (1.30)
600-650	61.11 (2.56)	54.58 (1.59)	48.70 (1.20)	45.71 (1.42)
650-700	64.82 (2.57)	56.46 (1.67)	50.15 (1.28)	45.93 (1.54)
700-750	67.08 (2.53)	57.70 (1.79)	50.74 (1.32)	45.42 (1.59)
750-800	68.16 (2.75)	58.29 (1.84)	50.83 (1.44)	44.93 (1.63)
800-850	68.31 (2.99)	58.82 (1.89)	51.11 (1.62)	44.90 (1.64)
850-900	69.24 (3.17)	58.31 (1.94)	51.60 (1.75)	45.14 (1.66)
900-950	67.49 (3.04)	57.95 (2.09)	52.13 (1.83)	45.62 (1.69)
950-1000	66.77 (3.04)	57.82 (2.20)	52.57 (1.88)	46.27 (1.76)
1000-1050	64.58 (3.11)	57.74 (2.27)	53.03 (1.93)	47.01 (1.85)
1050-1100	61.94 (3.25)	58.12 (2.34)	52.37 (1.99)	47.57 (1.96)
1100-1150	60.71 (2.45)	58.13 (2.44)	53.65 (2.06)	48.45 (2.08)
1150-1200	59.74 (3.65)	58.06 (2.53)	53.92 (2.16)	49.04 (2.20)

Table 3.5 Estimated mean labor hours used within consecutive 50,000 mile intervals. Standard errors are shown in parentheses.

Mileage interval (×1000)	Average miles per year 225000	250000	275000	300000
0-50	32.68 (.84)	31.85 (.89)	31.20 (.95)	31.52 (1.00)
50-100	33.15 (.85)	32.29 (.90)	31.78 (.95)	30.64 (1.00)
100-150	33.68 (.85)	32.62 (.90)	31.69 (.95)	30.70 (.99)
150-200	34.19 (.86)	32.86 (.91)	31.78 (.96)	30.77 (1.01)
200-250	34.66 (.89)	33.17 (.93)	31.89 (.97)	31.06 (1.04)
250-300	35.10 (.91)	33.33 (.95)	32.02 (1.01)	31.21 (1.06)
300-350	35.48 (.94)	33.40 (.98)	32.25 (1.06)	31.50 (1.11)
350-400	35.97 (.99)	33.66 (1.05)	32.59 (1.10)	31.78 (1.14)
400-450	36.34 (1.06)	34.43 (1.13)	33.34 (1.12)	31.78 (1.15)
450-500	37.20 (1.16)	34.75 (1.15)	33.52 (1.16)	32.65 (1.19)
500-550	38.33 (1.23)	35.30 (1.16)	33.81 (1.17)	32.57 (1.23)
550-600	39.16 (1.29)	35.76 (1.21)	34.12 (1.22)	32.99 (1.26)
600-650	39.61 (1.38)	36.07 (1.29)	34.39 (1.27)	32.72 (1.29)
650-700	39.91 (1.45)	36.03 (1.29)	34.53 (1.38)	33.18 (1.39)
700-750	40.16 (1.50)	36.03 (1.38)	34.25 (1.40)	33.02 (1.46)
750-800	40.46 (1.53)	36.33 (1.46)	33.77 (1.44)	33.18 (1.50)
800-850	41.22 (1.59)	36.70 (1.55)	33.87 (1.54)	31.98 (1.54)
850-900	41.54 (1.66)	37.42 (1.69)	33.71 (1.60)	32.18 (1.64)
900-950	41.97 (1.72)	37.98 (1.80)	33.55 (1.68)	32.08 (1.72)
950-1000	42.97 (1.83)	38.68 (1.91)	33.90 (1.83)	31.31 (1.77)
1000-1050	42.56 (1.89)	39.12 (2.00)	34.20 (1.97)	31.16 (1.86)
1050-1100	44.18 (1.99)	39.95 (2.12)	34.85 (2.09)	31.42 (1.91)
1100-1150	44.51 (2.10)	40.75 (2.21)	35.50 (2.21)	31.73 (2.06)
1150-1200	45.19 (2.21)	41.67 (2.24)	35.83 (2.38)	31.17 (2.16)

4 GROWTH CURVE ESTIMATION AND PREDICTION USING LINEAR MIXED MODELS

Introduction

A general model for the analysis of longitudinal data proposed by Harville [13] and Laird and Ware [17] can also be used for growth curve studies. In the previous two chapters attention has been focused on estimating the overall mean growth curve and using this curve to make predictions. The general model proposed by Harville and Laird and Ware, sometimes called the linear mixed model, can be used to predict individual growth curves as well as estimate the mean growth curve for the population. Predictions for a subject may then be obtained from that subject's individual growth curve.

The model

The linear mixed model for longitudinal data analysis is

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i \tag{4.1}$$

where y_i is an $n_i \times 1$ column vector of responses for unit i, X_i is an $n_i \times p$ design matrix, β is a $p \times 1$ vector of fixed coefficients, Z_i is an $n_i \times q$ design matrix for the random effects, γ_i, which are assumed to be independent with distribution $N(0, \sigma^2 B)$ where B is a general covariance matrix. The within unit errors, ϵ_i, are assumed to be independent with distribution $N(0, \sigma^2 W_i(\theta))$ and are also assumed to be independent of γ_i. This model allows for a different number of observations for each unit and irregular

inspection times. The covariance matrices $\sigma^2 W_i$ are allowed to vary from unit to unit but are parameterized by the vector θ. Let $V_i = Z_i B Z_i' + W_i$. Then the covariance matrix for the n_i observations taken on unit i is

$$\sigma^2 V_i = \sigma^2 (Z_i B Z_i' + W_i). \tag{4.2}$$

The parameterization of W_i will differ for different situations. For repeated measures a parameterization of $W_i(\theta)$ proposed by Jones and Ackerson [15] and Jones and Boadi-Boateng [16] allows for serial correlation in the within unit error structure. This parameterization makes sense for repeated measures and growth curve studies because it is reasonable to assume that errors within a unit could be correlated. When inspection times are irregular it is necessary to consider continuous time autoregressive processes. The simplest case is a continuous time AR(1) process.

A continuous time AR(1) process has correlation function

$$\rho(\tau) = e^{-\theta|\tau|}$$

where τ is the time interval between two successive observations. Under this assumption the within unit correlation matrix is

$$W_i = \begin{bmatrix} 1 & \cdot & \cdot & e^{-\theta|\tau|} \\ \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot \\ e^{-\theta|\tau|} & \cdot & \cdot & 1 \end{bmatrix}.$$

Estimation of the parameters

The log likelihood for this model multiplied by -2 is

$$l = \sum_{i=1}^{p} \left\{ n_i \log(2\pi\sigma^2) + \log|Z_i B Z_i' + W_i| + \frac{1}{\sigma^2}(y_i - X_i\beta)'(Z_i B Z_i' + W_i)^{-1}(y_i - X_i\beta) \right\}. \tag{4.3}$$

It is a function of β, σ^2, the parameters of the matrices W_i, and B. For a given B and θ, the weighted least squares estimate of β is

$$\hat{\beta} = \left\{\sum_{i=1}^{p} X_i'(Z_iBZ_i' + W_i)^{-1}X_i\right\}^{-1}\left\{\sum_{i=1}^{p} X_i'(Z_iBZ_i' + W_i)^{-1}y_i\right\} \tag{4.4}$$

and an estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{p}(y_i - X_i\hat{\beta})'(Z_iBZ_i' + W_i)^{-1}(y_i - X_i\hat{\beta}) \tag{4.5}$$

where $N = \sum_{i=1}^{p} n_i$ is the total number of observations from all the units. Substituting these estimates back into Equation 4.3 gives l concentrated with respect to β and σ^2.

$$l = N\log(2\pi\hat{\sigma}^2) + \sum_{i=1}^{p}\log|Z_iBZ_i' + W_i| + N \tag{4.6}$$

To constrain B to be nonnegative definite, B may be factored as

$$B = U'U \tag{4.7}$$

where U is an upper triangular matrix. Then Equation 4.6 becomes

$$l = N\log(2\pi\hat{\sigma}^2) + \sum_{i=1}^{p}\log|Z_iU'UZ_i' + W_i| + N. \tag{4.8}$$

Maximum likelihood estimates are obtained by minimizing Equation 4.8 with respect to θ, the vector which parameterizes W_i, and the elements of U. These estimates are then substituted back into Equations 4.4 and 4.5 to obtain estimates for β and σ^2.

To obtain a prediction for γ_i we note that for the i^{th} unit $Cov(\gamma_i, y_i) = \sigma^2BZ_i'$. Then, because of normality, we have

$$E(\gamma_i|y_i) = E(\gamma_i) + Cov(\gamma_i, y_i)[var(y_i)]^{-1}[y_i - E(y_i)] = BZ_i'V_i^{-1}(y_i - X_i\beta). \tag{4.9}$$

Assuming B and θ are known, a prediction for γ_i may be obtained by substituting $\hat{\beta}$ into Equation 4.9 resulting in

$$\hat{\gamma}_i = BZ_i'V_i^{-1}(y_i - X_i\hat{\beta}). \tag{4.10}$$

The prediction for γ_i is also empirical Bayes since it has the form $\hat{\gamma}_i = E(\gamma_i|y_i, \hat{\beta}, \theta)$. In practice B and θ are unknown, so their maximum likelihood estimates are substituted into Equation 4.10 to obtain

$$\hat{\gamma}_i = \hat{B}Z_i'\hat{V}_i^{-1}(y_i - X_i\hat{\beta}). \tag{4.11}$$

Standard errors for $\hat{\beta}$ and $\hat{\gamma}_i$

Since $\hat{\beta}$ and $\hat{\gamma}_i$ are both linear functions of y_i, their standard errors assuming B and θ are known may be written as

$$Var(\hat{\beta}) = \sigma^2 \left(\sum_{i=1}^{p} X_i'V_i^{-1}X_i \right)^{-1} \tag{4.12}$$

and

$$Var(\hat{\gamma}_i) = \sigma^2 BZ_i' \left\{ V_i^{-1} - V_i^{-1}X_i \left(\sum_{i=1}^{p} X_i'V_i^{-1}X_i \right)^{-1} X_i'V_i^{-1} \right\} Z_iB. \tag{4.13}$$

To assess the error of prediction the variation of γ_i and its correlation with $\hat{\gamma}_i$ must also be taken into account. Thus we have

$$Var(\hat{\gamma}_i - \gamma_i) = \sigma^2 \left(B - BZ_i' \left\{ V_i^{-1} - V_i^{-1}X_i \left(\sum_{i=1}^{p} X_i'V_i^{-1}X_i \right)^{-1} X_i'V_i^{-1} \right\} Z_iB \right). \tag{4.14}$$

When B and θ are unknown, estimates for the standard errors may be obtained by substituting the maximum likelihood estimates $\hat{B}$ and $\hat{V}_i^{-1}$ in place of B and V_i^{-1}.

Application to the truck data

We now consider the application of the linear mixed model to the truck maintenance data. Recall $y_{i,j}$ = cumulative labor hours and $t_{i,j}$ = cumulative mileage of the i^{th} truck at the end of the j^{th} year of service. We will consider hierarchical polynomial growth curve models to model the population mean growth curve and the individual growth curves for each truck. The use of polynomials to model growth curves is commonplace when the underlying model which generates the response variable is not well understood.

Rao [24] considered polynomials for the prediction of future observations in growth curve models for the special case with homogeneous inspection times.

Suppose the population mean growth curve is to be modeled by a k^{th} degree polynomial in t. The population design matrix for the i^{th} truck is of the form

$$X_i = \begin{bmatrix} 1 & t_{i,1} & t_{i,1}^2 & \cdots & t_{i,1}^{k-1} \\ 1 & t_{i,2} & t_{i,2}^2 & \cdots & t_{i,2}^{k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_{i,n_i} & t_{i,n_i}^2 & \cdots & t_{i,n_i}^{k-1} \end{bmatrix}.$$

In a hierarchical polynomial growth curve model, we allow some of the higher order degree terms to be common to all trucks. The polynomial terms which vary from truck to truck have a fixed part which is modeled by the matrix X_i and a random part which is modeled by the matrix Z_i. Thus for some $m < k$, the design matrix for Z_i is

$$Z_i = \begin{bmatrix} 1 & t_{i,1} & t_{i,1}^2 & \cdots & t_{i,1}^{m-1} \\ 1 & t_{i,2} & t_{i,2}^2 & \cdots & t_{i,2}^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_{i,n_i} & t_{i,n_i}^2 & \cdots & t_{i,n_i}^{m-1} \end{bmatrix}.$$

As an example suppose $k = 2$ and $m = 1$. Then the linear mixed model for the i^{th} truck becomes

$$y_i = \begin{bmatrix} 1 & t_{i,1} & t_{i,1}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{i,n_1} & t_{i,n_1}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & t_{i,1} \\ \vdots & \vdots \\ 1 & t_{i,n_1} \end{bmatrix} \begin{bmatrix} \gamma_{0,i} \\ \gamma_{1,i} \end{bmatrix} + \epsilon_i.$$

In this model the population mean growth curve is given by

$$y = \beta_0 + \beta_1 t + \beta_2 t^2$$

and the individual growth curve for the i^{th} truck is given by

$$y = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)t + \beta_2 t^2.$$

For this model, the constant and linear terms will vary from truck to truck but the quadratic term is common to all trucks.

Model selection

The problem of selecting a linear mixed model for the truck maintenance data will now be considered. When two models are fit to the same data using maximum likelihood and one model is a constrained version of the other, the likelihood ratio test can be used to test the null hypothesis that the model with more parameters is not a significantly better fit than the model with fewer parameters. In this application the parameters of interest are the elements of the vector β, the elements of the upper triangular matrix U, and θ, the scalar which parameterizes W_i. Let l_1 be -2 log likelihood for the constrained model and let l_2 be -2 log likelihood for the other model with r extra parameters. Under the null hypothesis that the r extra parameters are zero, we have

$$l_1 - l_2 \sim \chi^2_r. \tag{4.15}$$

A large value of $l_1 - l_2$ is evidence in favor of the alternate hypothesis that there is a significant improvement in the fit with the extra r parameters.

An approach to find the best linear mixed model for the truck maintenance data is to determine a good model for the fixed and random effects using the chi-square tests. Then AR(1) error structure can be introduced to see if there is a significant improvement in the fit. The value for -2 log likelihood for a series of linear mixed models is shown in Table 4.1. Note that if the order of the random part of the polynomial is m then there are $(m+1)(m+2)/2$ elements of the matrix U which are parameters in the log likelihood, Equation 4.8. For a Type 1 error rate of .01 the critical values from a chi-square distribution with 1 degree of freedom and 2 degrees of freedom are 6.635 and 9.210, respectively.

Table 4.1 -2 log likelihood for models fitted to the truck maintenance data

Order		Number of	
Fixed	Random	Parameters	-2 log likelihood
1	1	5	56662.27
2	1	6	55828.60
2	2	9	53964.80
3	1	7	55614.89
3	2	10	53939.13
3	3	14	53700.61
4	1	8	55610.61
4	2	11	53910.82
4	3	15	53685.78
4	4	20	53685.78
5	2	12	53907.04
5	3	16	53680.16
6	2	13	53903.89
6	3	17	53679.69

The model with a fourth degree polynomial in t for the fixed part and a cubic polynomial in t for the random part provides a good fit to the data. We will refer to this model as Model 430 to denote the fixed fourth degree polynomial in t and the random third degree polynomial in t and no autocorrelation among errors within units. Imposing AR(1) error structure on this model results in a -2 log likelihood value of 53683.57, so it appears that the AR(1) error structure is not needed after the random effects are included in the model.

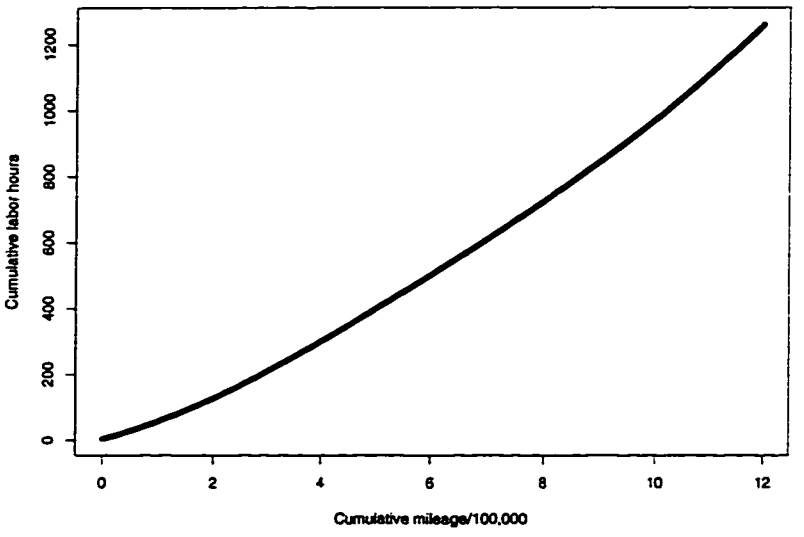

Figure 4.1 Estimated mean curve for Model 430

For the Model 430, $\theta = 0$, W_i is an identity matrix, and the parameter estimates are:

$$\hat{\beta} = \begin{bmatrix} 5.026 \\ 44.89 \\ 9.909 \\ -.7704 \\ .03003 \end{bmatrix} \qquad \hat{B} = \hat{U}'\hat{U} \qquad \hat{U} = \begin{bmatrix} .2847 & .2506 & .2114 & -.007927 \\ & 1.627 & -.2890 & .008451 \\ & & .3165 & -.03641 \\ & & & .004749 \end{bmatrix}$$

and $\hat{\sigma} = 18.35$.

The estimated mean curve is:

$$\hat{y} = 5.026 + 44.89t + 9.909t^2 - .7704t^3 + .03003t^4.$$

A graph of the estimated mean curve is shown in Figure 4.1

The individual growth curve for the i^{th} truck is given by

$$\begin{aligned} \hat{y}_i &= (5.026 + \hat{\gamma}_{0,i}) + (44.9 + \hat{\gamma}_{1,i})t + (9.909 + \hat{\gamma}_{2,i})t^2 \\ &\quad (-.7704 + \hat{\gamma}_{3,i})t^3 + .03003t^4 \end{aligned} \tag{4.16}$$

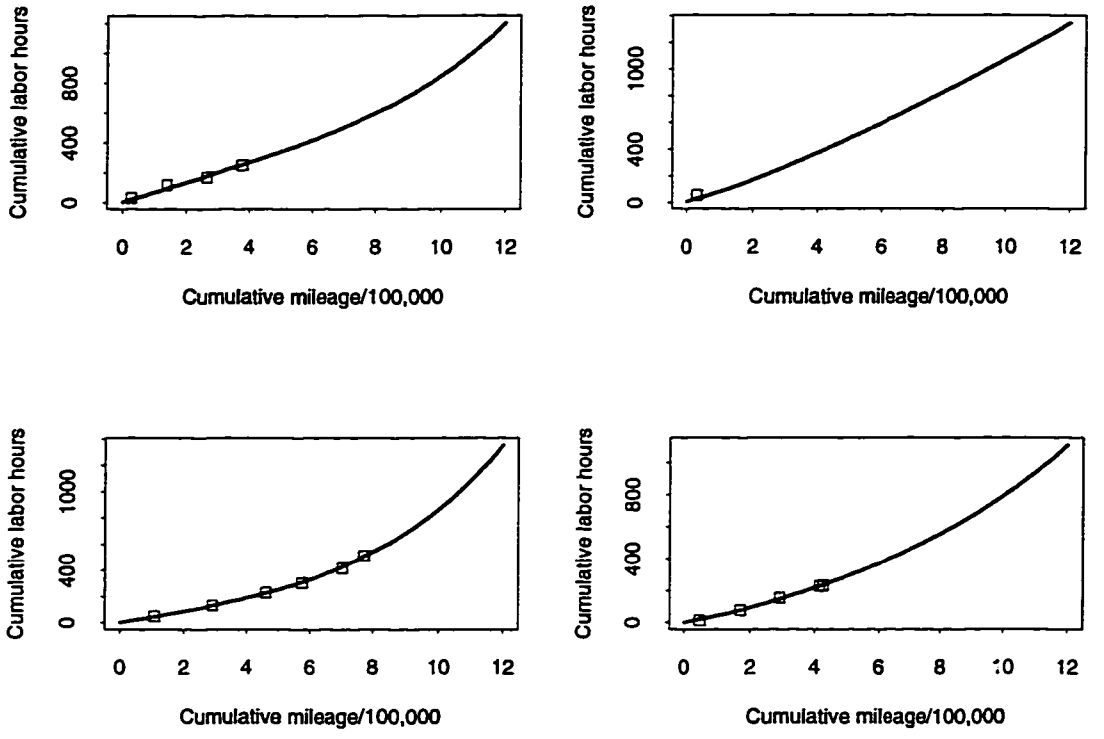

Figure 4.2 Individual growth curves for four trucks – Model 430

where $\hat{\gamma}_{0,i}$, $\hat{\gamma}_{1,i}$, $\hat{\gamma}_{2,i}$, and $\hat{\gamma}_{3,i}$ are the predicted random effects for the i^{th} truck. Graphs of the individual growth curves and observed data for four individual trucks are shown in Figure 4.2. Note that the fit of the individual curves to the observed data appears to be quite good. Also note that the truck with only one observation has an estimated individual growth curve. Thus, all four random effect terms can be predicted even when there are fewer than five observations per truck. The sums of squares associated with the fit of the model to the complete data set are shown in Table 4.2.

In Chapter 3 the curve fitting procedure *loess* was used to estimate the mean growth curve. Figure 4.3 is a plot the mean growth curves estimated by the mixed model (Model 430) and by loess. The two estimated curves agree quite well for cumulative mileages below 600,000 miles. The mean curve estimated by Model 430 rises more rapidly than the loess estimated mean curve for cumulative mileages above 600,000. Apparently, loess

Table 4.2 Sums of squares Model 430

Source	Sum of squares
Model	262,292,646
Error	1,103,108
Corrected total	263,395,754
R^2=.9958	

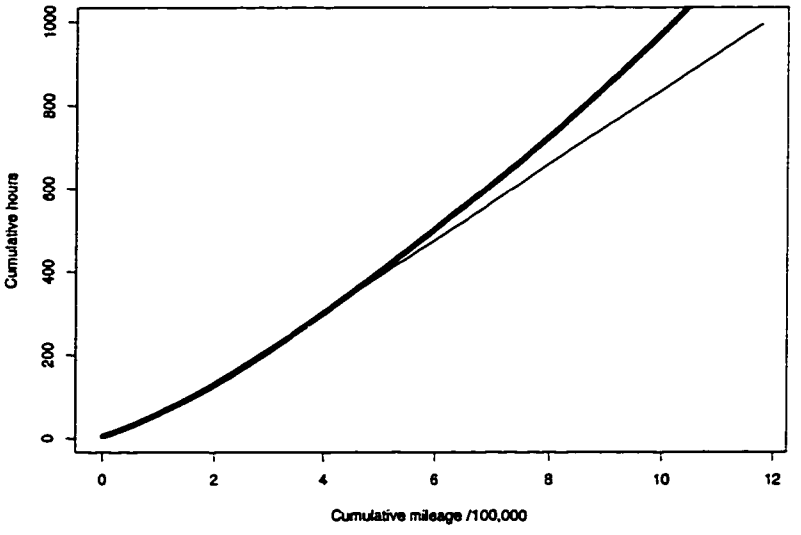

Figure 4.3 Estimated mean growth curve. Model 430 (dark curve) and loess estimated curve (light curve)

is more heavily influenced than the mixed model by high mileage trucks with relatively low labor hour requirements. The proportion of high mileage trucks with low labor hour requirements is relatively small for the truck data, so the mixed model allows these trucks to deviate from the estimated mean curve via the random effects. The mixed model estimated mean curve is determined by the large proportion of trucks with lower cumulative mileages. In the truck study, 75% of the 1182 units had cumulative mileages less than 675,000 miles.

Using the model for prediction of future values

The estimation of values on the population mean curve is no different from ordinary weighted least squares regression. Let t denote a possible value of the cumulative mileage variable t. For a fourth degree polynomial we would have

$$x = [\ 1 \quad t \quad t^2 \quad t^3 \quad t^4\ ].$$

Then, the estimated population mean for a given x is

$$\hat{y} = x\hat{\beta}$$

which has variance

$$\text{var}(\hat{y}) = \sigma^2 x(\sum X_i' V_i^{-1} X_i)^{-1} x' = x\text{cov}(\hat{\beta})x'.$$

With the linear mixed model it is possible to make predictions for individual trucks. Suppose we wish to predict the cumulative labor hours for a truck at future cumulative mileage, t_p. For a model with a fourth degree polynomial for fixed effects and a third degree polynomial for random effects we would have

$$x_p = [\ 1 \quad t_p \quad t_p^2 \quad t_p^3 \quad t_p^4\ ]$$

$$z_p = [\ 1 \quad t_p \quad t_p^2 \quad t_p^3\ ].$$

First consider the case of a truck for which there is no previous data. Since there is no prediction for γ_i for this truck the prediction of the cumulative labor hours is given by the population mean curve

$$\hat{y}_p = x_p\hat{\beta}. \tag{4.17}$$

But the variance of the prediction is given by

$$\text{var}(y - \hat{y}_p) = x_p\text{cov}(\hat{\beta})x_p' + \sigma^2 z_p B z_p' + \sigma^2. \tag{4.18}$$

Now consider the case where the i^{th} truck has previous data. The model for this truck may be written as:

$$\begin{bmatrix} y_i \\ y_p \end{bmatrix} = \begin{bmatrix} X_i \\ x_p \end{bmatrix} \beta + \begin{bmatrix} Z_i \\ z_p \end{bmatrix} \gamma_i + \begin{bmatrix} \epsilon_i \\ \epsilon_p \end{bmatrix}. \tag{4.19}$$

Then

$$\text{Var} \begin{bmatrix} y_i \\ y_p \end{bmatrix} = \sigma^2 \begin{bmatrix} Z_i B Z_i' + W_i & Z_i B z_p' \\ z_p B Z_i' & z_p B z_p' + 1 \end{bmatrix}. \tag{4.20}$$

Assuming β, B, $\hat{\sigma}^2$ are known, the minimum mean squared error predictor of y_p based on y_i is the conditional mean

$$\hat{y}_{p,B,\beta} = E(y_p|y_i) = x_p\beta + z_p B Z_i' V_i^{-1}(y_i - X_i\beta) \tag{4.21}$$

where $V_i = Z_i B Z_i' + W_i$. An empirical Bayes estimator for y_p may be obtained by substituting $\hat{\beta}$ into Equation 4.21 resulting in

$$\hat{y}_{p,B} = x_p\hat{\beta} + z_p B Z_i' V_i^{-1}(y_i - X_i\hat{\beta}) \tag{4.22}$$

which has variance

$$\text{var}(y - \hat{y}_{p,B}) = x_p \text{cov}(\hat{\beta}) x_p' + z_p \text{cov}(\hat{\gamma}_i - \gamma_i) z_p' + \sigma^2. \tag{4.23}$$

When B is unknown, the maximum likelihood estimate $\hat{B}$ may be substituted to obtain

$$\begin{aligned} \hat{y}_p &= x_p\hat{\beta} + z_p \hat{B} Z_i' V_i^{-1}(y_i - X_i\hat{\beta}) \\ &= x_p\hat{\beta} + z_p\hat{\gamma}_i. \end{aligned} \tag{4.24}$$

These estimators and predictors were derived by Harville [13] and considered by Laird and Ware [17] among others. Rao [24] [23] and Reinsel [25] have used these estimators in the context of predicting future observations in growth curve models.

Interpretation of $\hat{\gamma}_i$ and $\hat{y}_p$

Suppose the linear mixed model is written as

$$(y_i - X_i\beta) = Z_i\gamma_i + \epsilon_i$$

and Z_i is assumed to be full rank. Rao [23] has shown that an expression for $\hat{\gamma}_i$ equivalent to Equation 4.10 is given by

$$\hat{\gamma}_i = \hat{\gamma}_i^{(l)} - \sigma^2(Z_i'Z_i)^{-1}(B + (Z_i'Z_i)^{-1})^{-1}(\hat{\gamma}_i^{(l)}) \tag{4.25}$$

where $\hat{\gamma}_i^{(l)} = (Z_i'Z_i)^{-1}Z_i'(y_i - X_i\beta)$ is the least squares estimator of γ_i obtained by treating γ_i as a fixed effect in the regression of $(y_i - X_i\beta)$ on Z_i. Written in this form we see that $\hat{\gamma}_i$ is a weighted linear combination of $\hat{\gamma}_i^{(l)}$ and the assumed prior mean of γ_i, i.e. $E(\gamma_i) = 0$. Rao [24] proposed this estimator in his paper on predicting future observations in growth curve models.

There is some confusion in the literature regarding the interpretation of $\hat{y}_p$ and $\hat{\gamma}_i$ in growth curve models. Note that for the truck data we fit a third order polynomial model as the within-unit model even though many trucks have fewer than four observations. For these trucks Z_i does not have full column rank and $Z_i'Z_i$ is not positive definite. Therefore the interpretation of $\hat{\gamma}_i$ afforded by Equation 4.25 does not apply since $\hat{\gamma}_i^{(l)}$ is not even estimable. Laird and Ware [17] consider the estimator in Equation 4.10 and also state that $\hat{\gamma}_i$ is a weighted linear combination of $\hat{\gamma}_i^{(l)}$ and 0, but they also fail to note that this is true only when Z_i has full column rank.

Jones and Boadi-Boateng [16] analyze a data set with a linear mixed model with a first order polynomial for the overall mean curve (fixed effects) and a first order polynomial for the within subject model (random effects). They note that many subjects have only a single observation and yet straight lines are being fitted as the within subject model. They go on to state that the estimated intercept and slope for a subject are shrinkage estimates that are shrunk toward the overall mean line, and if a subject has

only a single observation the estimated slope for that subject will be the same as the slope for the overall mean line, but the estimated line will not go through the single observation; it will be shifted toward the overall mean line.

This interpretation by Jones and Boadi-Boateng is in error. Both the estimated intercept and the estimated slope for a subject with only one observation will be different from the overall mean line as an inspection of Equation 4.10 will show. Furthermore, since Z_i is not full column rank in this case, one cannot talk about the slope and intercept for the subject as being weighted linear combinations of 0 and $\hat{\gamma}_i^{(l)}$ since $\hat{\gamma}_i^{(l)}$ is not estimable. For a subject with only one observation what we can say is that

$$\hat{y}_i = X_i\hat{\beta} + Z_i B Z_i' V_i^{-1}(y_i - X_i\hat{\beta}) \tag{4.26}$$

and therefore the predicted response for a subject with one observation at X_i is shrunk from the observed value y_i toward the estimated group mean $X_i\hat{\beta}$.

Choosing a model via cross validation

The intended use of the linear mixed model as applied to the truck data is to predict cumulative labor hours at a future cumulative mileage, t_p. The model selection procedure presented earlier using chi-square tests assesses the quality of fit of the different models. A model selected by this procedure may fit the data well but may not be the best model to use for prediction. The model selected by the chi-square test procedure called for a fourth order polynomial for the between unit model and a third order polynomial for the within unit model. Is this the best model for prediction or would a lower order model do as well or better?

One possible way to assess the predictive ability of a model is to use cross validation. As in Chapter 2, the data are split into a fitting sample and a validation sample. The candidate models are fitted using the fitting sample and then the resulting models are

Table 4.3 Cross validation results

Order		
Fixed	Random	$\frac{1}{939}\sum(y-\hat{y})^2$
2	1	2581.68
3	1	2446.93
3	2	1998.56
4	2	1988.47
4	3	1822.51

used to estimate the responses in the validation sample. Candidate models may then be compared by using the average of the sum of the squared residuals.

The validation sample we selected for the truck data was the last observation for each truck that was still in service at the end of 1993. Some trucks in the data set were removed from service prior to 1993 so it is realistic to assume that future predictions for these trucks are not desired. Of the 5344 total observations in the truck data set this data splitting scheme placed 939 observations in the validation sample and 4405 observations in the fitting sample. Results for a series of linear mixed models is shown in Table 4.3.

The results of Table 4.3 show that the lower order models do not predict 1993 cumulative labor hours as well as the model selected by the chi-square tests.

Checking model assumptions

We now check the distributional assumptions for the random coefficents, γ_i, and the within unit errors, ϵ_i. In Equation 4.1 we assumed that the random effects, γ_i, had distribution $N(0, \sigma^2 B)$ where B is a general covariance matrix and that the within unit errors, ϵ_i, had distribution $N(0, \sigma^2 W_i(\theta))$. We later found that there appeared to be no autocorrelation of errors within units and so the assumption for the within unit errors became $\epsilon_i \sim N(0, \sigma^2 I)$.

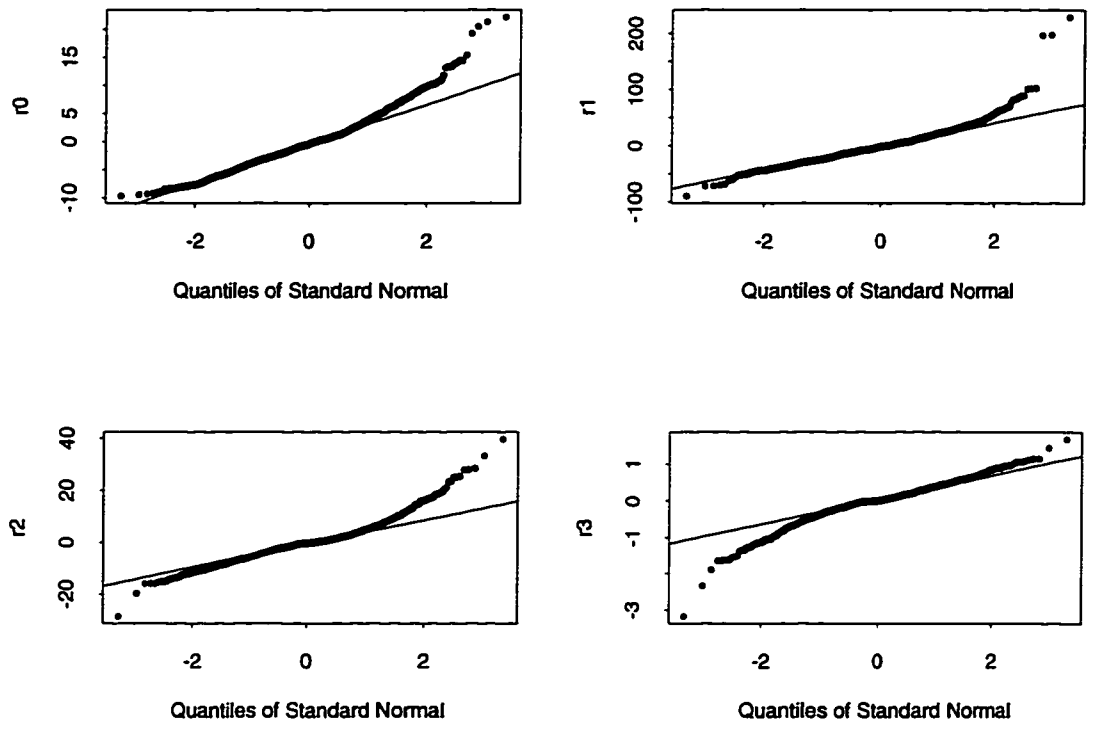

Figure 4.4 Normal probability plots for random coefficients. Clockwise from upper left: intercept, linear, cubic, quadratic

Normal probability plots for the estimated random coefficients are displayed in Figure 4.4. The plot labeled $r0$ on the vertical axis is the plot for random intercept terms, the plot labeled $r1$ is for the random linear terms, and so on. A line through the lower and upper quartiles is shown on each plot so that one can judge the straightness of the points. It appears from the plots that the distributions for the predicted random intercept, linear, and quadratic coefficents may be skewed right and the distribution for the predicted random cubic coefficient may be skewed left, but that a gross departure from normality does not appear to be present.

Figure 4.5 is a normal probability plot of the residuals, $\hat{\epsilon}_i$. The distribution appears to be heavy tailed but otherwise symmetric.

The constant variance assumption of the residuals was checked by partitioning the residuals into six classes according to cumulative mileage. The classes were 0-200,000

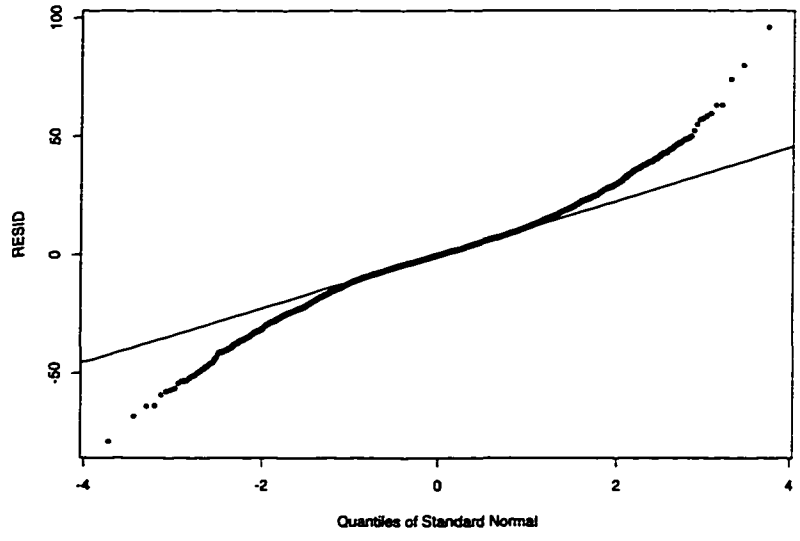

Figure 4.5 Normal probability plot of residuals

Table 4.4 Variance of residuals within cumulative mileage classes

Cumulative mileage	Sample variance	Number of residuals
0-200,000	186.35	1716
200,000-400,000	201.10	1541
400,000-600,000	239.33	1240
600,000-800,000	185.56	583
800,000-1,000,000	265.84	187
1,000,000-1,200,000	253.26	77

miles, 200,000-400,000 miles, and so on. The sample variance of the residuals within each class were computed and are presented in Table 4.4 The variance of the residuals increases slightly for increasing cumulative mileage, but does not appear to present a problem.

Incorporating pace into the mixed model analysis

The variable pace, the average number of miles a truck is driven, was introduced in Chapters 2 and 3 and was found to improve estimates of cumulative labor hours. In this section we incorporate pace as a covariate into the mixed model analysis.

The best model found using only cumulative mileage, t, as the explanatory variable was a fixed fourth order polynomial in t with a random third order polynomial in t. This model was called Model 430. We will incorporate pace into this model. The question now arises as to which design matrix, X_i or Z_i (or both), should contain the variable pace. It is not clear whether pace can be regarded as a random effect, but a good argument can be make for keeping pace in X_i, the fixed effect design matrix.

1. If the fixed effect part of the model can be improved, the need for more random terms is reduced.

2. If the model contains many random terms, the computer program which estimates the parameters of the model has difficulty in converging to a solution.

Thus, pace will be incorporated by interacting pace with every fixed effect term in Model 430. Let r=pace. The model for the i^{th} truck is

$$\begin{aligned} y_i &= \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 \\ &\quad + \beta_5 r + \beta_6 rt + \beta_7 rt^2 + \beta_8 rt^3 + \beta_9 rt^4 \\ &\quad + \gamma_{0,i} + \gamma_{1,i} t + \gamma_{2,i} t^2 + \gamma_{3,i} t^3 + \epsilon_i \end{aligned} \tag{4.27}$$

Fitting the above model to the data gave a -2 log likelihood of 53440.71. Compared to a value of 53685.78 for the model without the five pace terms, there is a decrease of 245.07 in -2 log likelihood. The .01 critical value of a χ^2 distribution with 5 degrees of freedom is 15.09, so at least one of the five pace terms is significant.

Parameter estimates for the fitted model are:

$$\hat{\beta} = \begin{bmatrix} 12.08 \\ 33.35 \\ 25.65 \\ -2.843 \\ .1504 \\ -5.231 \\ 7.695 \\ -11.45 \\ 1.350 \\ -.06965 \end{bmatrix} \qquad \hat{B} = \hat{U}'\hat{U} \qquad \hat{U} = \begin{bmatrix} .2611 & .2253 & .1905 & -.01037 \\ & 1.622 & -.2710 & .008729 \\ & & .2988 & -.03288 \\ & & & -.01138 \end{bmatrix}$$

$\hat{\sigma} = 18.41$.

We have also considered using cross-validation to choose a model. Using the same data splitting scheme as before, the average of the sum of the squared residuals was 1701.81. This compares to a value of 1822.51 (Table 4.3) for the model without the pace terms. Introduction of pace appears to have improved the predictive ability of the model.

A perspective plot of the estimated mean surface is shown in Figure 4.6. A comparison with the loess mean surface (incorporating pace) and the mixed model mean surface for the pace values 50,000, 100,000, 200,000 and 300,000 average miles per year is shown in Figure 4.7. The mixed model curves are the lighter curves. For the pace values 50,000 and 100,000 the agreement between the curves is good for cumulative mileages less than 800,000. Above 800,00 cumulative miles, the mixed model curve is greater than the loess curve. At the pace value of 200,000 the loess estimated curve is greater than the mixed model curve for all cumulative mileages. The pace value of 300,000 represents an

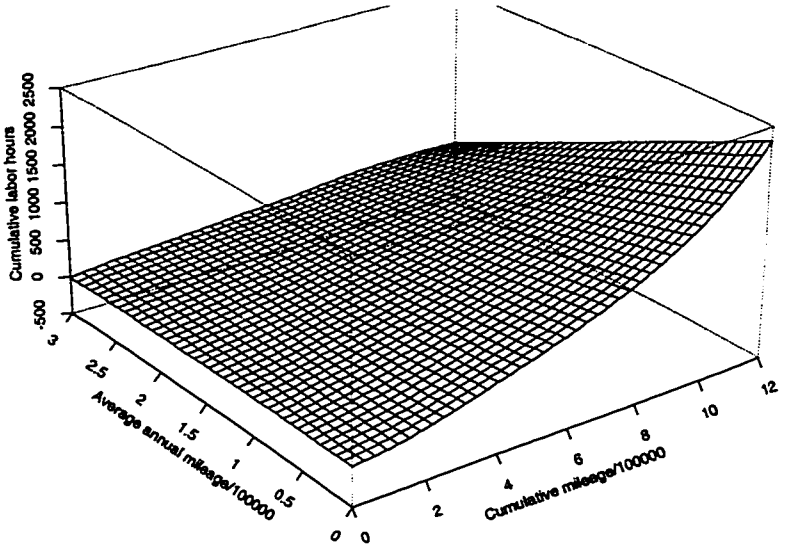

Figure 4.6 Cumulative labor hours versus cumulative mileage and pace – mixed model

extreme extrapolation for the model, but the mixed model curve does exhibit a decrease in cumulative labor hours for cumulative mileage above 1,000,000 miles.

Discussion and conclusion

This chapter has considered the linear mixed model for estimating and predicting cumulative labor hours. Some comparisons with the linear interpolation approach in Chapter 2 and loess in Chapter 3 can be made.

1. The linear mixed model provides estimates of the overall mean growth curve and growth curves for individual trucks.

2. Implementing the linear mixed model is more complicated than loess or the interpolation method. The major problem are proper choices for the between unit model (the overall mean curve) and the within unit model (the individual growth curve).

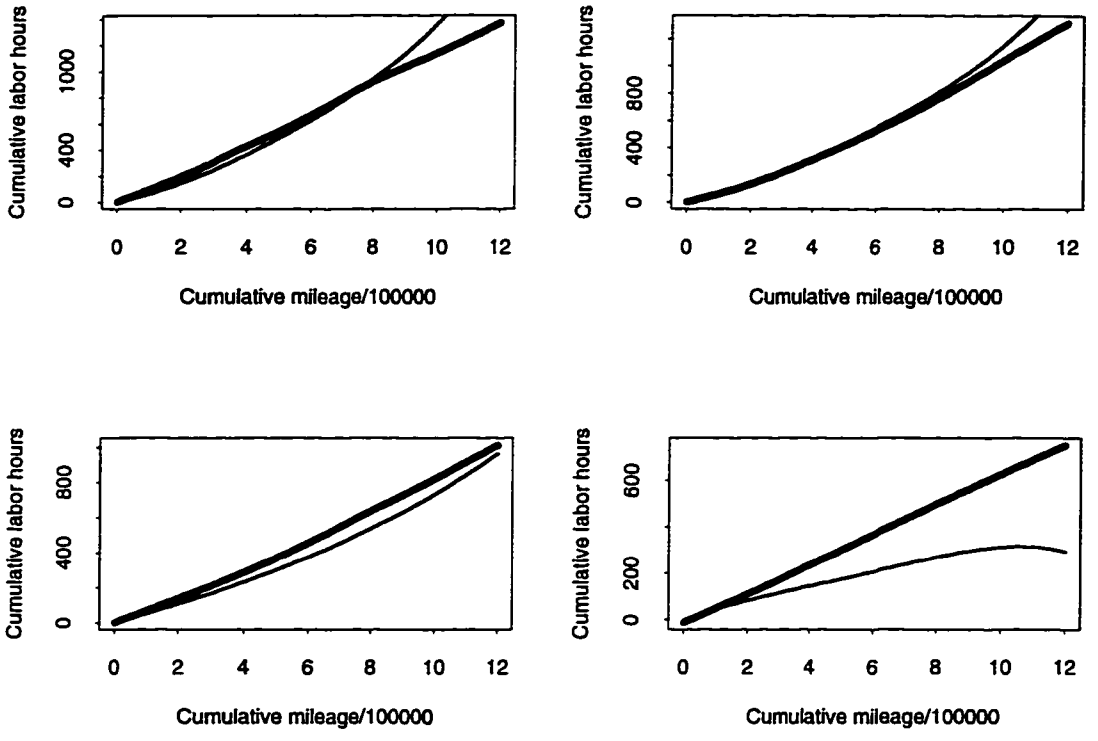

Figure 4.7 Loess mean curve (darker curve) and mixed model mean curve. Pace values from upper left are 50,000, 100,000, 200,000, and 300,000 miles per year

3. The linear mixed model allows one to extrapolate the estimated population mean curve and an individual's predicted growth curve.

4. The computational algorithms for calculating parameter estimates and predictions are more numerically intensive than the other two procedures. However bootstrap methods to calculate standard errors were not needed as was the case for loess.

5 A COMPARISON OF METHODS USING 1994 DATA

Introduction

Three methods have been presented in previous chapters for the analysis of the truck maintenance data. A nonparametric interpolation method was presented in Chapter 2. Locally weighted regression regression, or *loess*, was presented in Chapter 3. Linear mixed models were used in Chapter 4. Estimation of parameters and model selection were performed using 1993 data in all three chapters. In this chapter we will use models fitted to 1993 data to predict results for 1994. The actual 1994 data will then be compared to the predicted results.

Methods used

The 1994 data set has 1305 units with a total of 6069 observations. Of the 1305 units, there were 731 units which were still in service at the end of 1994. The cumulative mileage and cumulative labor hours at the end of 1994 for these 731 units are the data to which the three data analysis methods will be applied.

The comparison of the three methods was carried out by predicting the *increase* in cumulative labor hours in 1994 using information on the cumulative mileage at the end of 1993 and the actual mileage during 1994. Prediction errors were calculated as the observed increase in labor hours minus the predicted increase in labor hours. Performance of the three methods may then be compared using the average prediction error and the average squared prediction error.

Table 5.1 Average squared prediction error of the three methods

	Linear interpolation	Loess	Mixed model
Average prediction error	3.04	0.01	4.88
Average squared prediction error	1319.77	1373.13	1124.22

The models used in this comparison were the best models for the respective methods as determined in the earlier chapters.

1. The linear interpolation method with stratification with respect to pace. The stratification scheme with 8 pace categories was used. The categories (in thousands of miles per year) were 0-30, 30-50, 50-90, 90-130, 130-170, 170-210, 210-230, > 230.

2. Loess. A loess model incorporating pace and with specifications $\lambda = 1$ and $\alpha = 0.25$ was used. Weights as determined in Chapter 3 were used to handle the nonconstant variance of the errors.

3. Mixed Model. Model 430 with pace was used.

Results for the three methods are shown in Table 5.1. The mixed model had the highest average prediction error and the lowest average squared prediction error. Loess produced the smallest average prediction error. Average squared prediction error was similar for linear interpolation and for loess.

A further comparison of the three methods may be made by subdividing the 731 observations according to pace or by subdividing the observations according to cumulative mileage at the end of 1994. Tables 5.2 and 5.3 show the results when the data are subdivided according to the stratification scheme used for the linear interpolation method. Table 5.2 shows that the mixed model had positive average prediction errors for all pace categories except the last one. Linear interpolation and loess had positive average prediction errors in the middle pace categories and negative average prediction errors in the lower and higher pace categories. Table 5.3 shows that the mixed model had

Table 5.2 Average prediction error of the three methods by pace

Pace (thousands of miles per year)	Number of observations	Linear interpolation	Loess	Mixed model
0-30	9	-10.92	-3.27	7.92
30-50	19	-10.57	-13.42	6.08
50-90	168	4.18	3.39	10.30
90-130	343	4.79	1.95	2.91
130-170	104	3.98	-5.04	2.70
170-210	82	-1.53	-3.46	4.13
210-230	5	-17.71	-18.75	8.56
>230	1	-25.90	-46.57	-11.27

Table 5.3 Average squared prediction error of the three methods by pace

Pace (thousands of miles per year)	Number of observations	Linear interpolation	Loess	Mixed model
0-30	9	688.28	467.24	398.72
30-50	19	265.30	289.76	137.16
50-90	168	975.17	1065.78	972.85
90-130	343	1217.35	1268.64	1032.69
130-170	104	2544.54	2662.81	2180.29
170-210	82	1210.09	1147.46	848.57
210-230	5	1520.90	1331.93	300.22
>230	1	670.68	2168.57	127.16

smaller average squared prediction errors than the other two methods in all pace categories. The mixed model performed especially well in the highest three pace categories.

Tables 5.4 and 5.5 show the results when the data are subdivided according to cumulative mileage at the end of 1994. The results of Table 5.4 for average prediction error do not reveal any discernable pattern except for the large average prediction errors in the 400,00-600,000 mile class produced by all three methods. Table 5.5 shows that the mixed model had the lowest average squared prediction errors for all cumulative mileage classes.

Figures 5.3, 5.1, and 5.2 are plots of prediction errors versus cumulative mileage for

Table 5.4 Average prediction error of the three methods by 1994 cumulative mileage

Cumulative mileage	Number of observations	Linear interpolation	Loess	Mixed model
0-200,000	205	-3.49	-6.76	3.54
200,000-400,000	146	0.59	1.09	2.81
400,000-600,000	124	17.38	19.67	13.93
600,000-800,000	199	0.63	-7.36	1.56
800,000-1,000,000	50	9.61	3.70	6.62
1,000,000-1,200,000	7	12.41	10.45	8.29

Table 5.5 Average squared prediction error of the three methods by 1994 cumulative mileage

Cumulative mileage	Number of observations	Linear interpolation	Loess	Mixed model
0-200,000	205	460.89	468.98	448.93
200,000-400,000	146	1454.73	1403.78	717.76
400,000-600,000	124	1920.91	1965.92	1912.03
600,000-800,000	199	1583.86	1779.20	1459.52
800,000-1,000,000	50	2054.50	2060.96	1932.62
1,000,000-1,200,000	7	252.97	254.47	105.24

the three methods. For all three methods there was a set of five similar trucks that produced unusually large positive residuals. These trucks had 1994 cumulative mileages between 730,000 and 810,000 miles. All five were classified in the 130,000-170,000 average miles per year pace category, travelled approximately 80,000 miles in 1994, and required 200-250 labor hours for 1994. Although the data for these trucks do not appear to be in error, it is noteworthy that removing them produces average squared prediction errors in the 130,000-170,000 pace category of 1555.92, 1908.96, and 1325.75 for linear interpolation, loess, and the mixed model, respectively (compare with Table 5.3, 130-170 pace class).

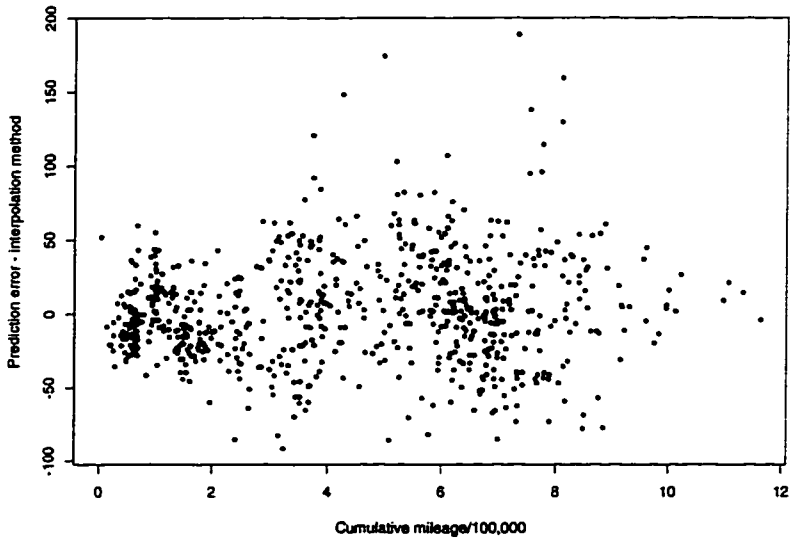

Figure 5.1 Prediction errors versus cumulative mileage - linear interpolation

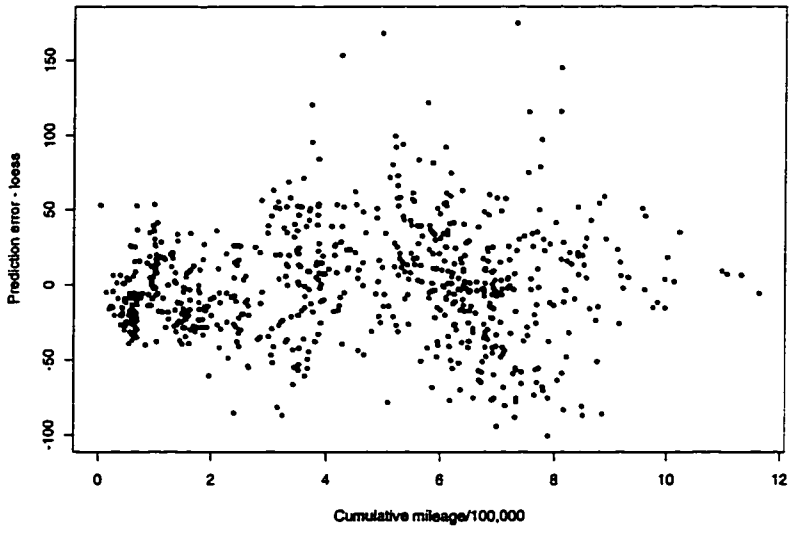

Figure 5.2 Prediction errors versus cumulative mileage - loess

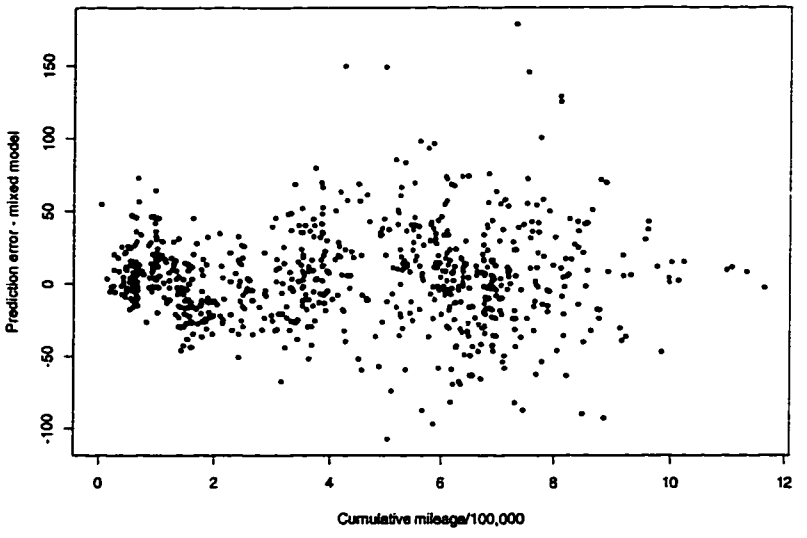

Figure 5.3 Prediction errors versus cumulative mileage - mixed model

Conclusion

The application of the three methods to the 1994 data shows that the mixed model generally outperforms the other two methods when average squared prediction error is the criterion used to compare the methods. Linear interpolation and loess are better when average prediction error is the criterion used for comparison. The small improvement in average squared prediction error gained by using the mixed model is partially offset by its computational demands. Also, a new model must be fitted at the end of each year when new data become available. Loess is also computationally intensive although it is available on S-Plus. The computational simplicity of the linear interpolation method is very appealing, especially if the method is adopted by a business. As noted in Chapter 2, the linear interpolation method can be implemented in a spreadsheet program or in SAS using data statements and PROC MEANS.

As noted in Chapter 1, the trucking firm has several maintenance terminals located across the United States. Maintenance for a given truck is generally performed at the same terminal for the duration of the lease. An analysis comparing the performance of

the maintenance terminals can be done by classifying the trucks according to maintenance terminal.

Strategies for improving the accuracy and variance of the predictions is a topic for future research. One possible strategy is to delete older data from the data set. Improvements in the construction of trucks and in maintenance techniques are not reflected in the older data. Some trucks have been removed from service prior to 1993, so the entire history of these trucks might be deleted. Trucks still in service at the end of 1993 could have part of their early history deleted.

Another strategy for improving predictions applies only to the mixed model. Using the mixed model to predict a future observation for an individual is an extrapolation of that individual's growth curve. In a longitudinal study with irregular inspection times it is possible that this same future observation is not an extrapolation for the overall mean curve. The usual approach for predicting the future observation is to simply use the individual's growth curve. Improvements in prediction could result if this prediction were "shrunk" toward the overall mean curve. Assuming this is the case, the appropriate amount of shrinkage needed would be a topic for future research.

APPENDIX A FORTRAN PROGRAM FOR INTERPOLATION

This is a Fortran program for computing estimated increases in cumulative labor hours over 50,000 mile intervals.

```
c
c  This program computes the estimated increase over 50000 mile
c  intervals
c
      integer nobs(5000),milecat(5000)
      integer ns,nt,ii,milec,unit
      real y(10000),t(10000),res(12),pred(12)
      character*32 datafil,filnam
      common ns,nobs,y,t,nt,milecat,milec
      ns=0
      nt=0
      ii=1
      write(*,*)'enter the name of data for interpolation alg'
      read(*,' (a)')datafil
      open(1,file=datafil,status='old')
      read(1,*)ncv,ntvcv
      do 5 i=1,10000
         y(i)=0.0
         t(i)=0.0
 5    continue
 10   continue
      read(1,*,end=21)  nobs(ii),milecat(ii),(res(i),i=1,nobs(ii))
      read(1,*,end=21)  (pred(i),i=1,nobs(ii))
      do 20 i=1,nobs(ii)
         t(nt+i+1)=res(i)
         y(nt+i+1)=pred(i)
```

```
 20   continue
      nt=nt+nobs(ii)+1
      ns=ns+1
      ii=ii+1
      go to 10
 21   continue
      jj=1
      write(*,*)'enter the name of the output file'
      read(*,' (a)')filnam
      open(2,file=filnam,status='new')
      do 35 jj=1,24
            x2=.5*jj
            x1=x2-.5
       do 30 i=1,10
         milec=i

      call interpolate(x1,x2,pdnd,p1,p1sq,p2,p2sq,p1p2,j)
c      write(2,*)unit,x1,x2,pdnd
c      write(2,*)p2sq,p1sq,p2,p1,p1p2,j
      if (j.gt.1) then
            call stderr(s,p1,p1sq,p2,p2sq,p1p2,j)
            else
               s=0
            endif
      write(2,50)milec,x1,x2,pdnd,s
 50    format(1x,i4,3x,4g11.4)
 30   continue
 35   continue
      stop
      end
c
c
c
      subroutine interpolate(x1,x2,pdnd,p1,p1sq,p2,p2sq,p1p2,j)
      integer nt,j,ns,nlp,milec
      integer nobs(5000),milecat(5000)
      real y(10000),t(10000)
      common ns,nobs,y,t,nt,milecat,milec
      nlp=0
      pdnd=0
      j=0
      p1=0
      p2=0
      p1sq=0
```

```
      p2sq=0
      p1p2=0
      do 1000 ii=1,ns
         k=nobs(ii)
         if (milecat(ii).ne.milec)  goto 2500
         do 2000 i=1,k
           n1=nlp+i
           n2=nlp+i+1
              if (t(n1).le.x2 .and. x2.le.t(n2)) then
            a=(((y(n2)-y(n1))*(x2-t(n1)))/(t(n2)-t(n1)))+y(n1)
            p2=p2+a
            p2sq=p2sq+(a**2)
            j=j+1
            do 3000 l=1,k
               m1=nlp+l
               m2=nlp+l+1
                   if (t(m1).le.x1 .and. x1.le.t(m2)) then
               b=(((y(m2)-y(m1))*(x1-t(m1)))/(t(m2)-t(m1)))+y(m1)
               p1=p1+b

               p1sq=p1sq+(b**2)
               p1p2=p1p2+(a*b)
               goto 2500
               endif
 3000        continue
           endif
 2000    continue
 2500    nlp=nlp+nobs(ii)+1
 1000    continue
       if (j.ne.0) then
           pdnd=(p2/j)-(p1/j)
         else
            pdnd=0.0
         endif
       return
       end
c
c
      subroutine stderr(s,p1,p1sq,p2,p2sq,p1p2,j)
      ssp2=p2sq-((p2**2)/j)
      ssp1=p1sq-((p1**2)/j)
      ssp1p2=p1p2-((p1*p2)/j)
      var=(ssp2+ssp1-(2*ssp1p2))/(j-1)
      if (var.le.0) then
```

```
          var=0.0
          endif
c       write(2,*)var
        s=sqrt(var/j)
        return
        end
```

APPENDIX B S-PLUS COMMANDS FOR M-PLOT

These are the S-Plus commands used to generate an M-plot. The truck data were stored in the S-Plus file cot93.

```
Commands for generating data for M-plot
using h=0.05


attach(cot93)

cot.m.05<-loess(Clhour ~ Ltdmile, cot93, weights=wts,
degree=1, span=0.05)

M<-list()      [Mean squared error]

B<-list()      [Bias]

V<-list()      [contribution of variance
                i.e. equivalent number of
                parameters]


> cot.m.05
Call:
loess(formula = Clhour ~ Ltdmile, data = cot93,
weights = wts, span = 0.05, degree = 1)

 Number of Observations:          5344
 Equivalent Number of Parameters: 37.1
 Residual Standard Error:         104.2
 Multiple R-squared:              0.5
```

```
 Residuals:
    min  1st Q median 3rd Q   max
 -392.5 -45.69 -4.828 37.37 574.3


[so, sigma_h^2  is 104.1577**2 or 10848.83]

cot.m.05$inference$s = Residual Standard Error
cot.m.05$surface$enp = Equivalent Number of
                       Parameters


 for(i in seq(16)){
   s<-(i*.05)
   print(s)
   cot.m.05<-loess(Clhour ~ Ltdmile, cot93,
             weights=wts, degree=1, span=s)
   RSE<-cot.m.05$inference$s
   enp<-cot.m.05$surface$enp
   sse<-sum((residuals(cot.m.05)*sqrt(wts))**2)
   b<-(sse/10848.83)-(sse/(RSE**2))
   M3[[i]]<-b+enp
   V3[[i]]<-enp
   }

---------------------------

Commands for generating an M-plot with
Ltdmile and Milerate as the explanatory
variables.


cot.m.05<-loess(Clhour ~ Ltdmile*Milerate, cot93,
weights=wts.pace, degree=1, span=0.05)

M1<-list()      [Mean squared error]

B1<-list()      [Bias]

V1<-list()      [contribution of variance
                i.e. equivalent number of
                parameters]

 cot.m.05
```

```
Call:
loess(formula = Clhour ~ Ltdmile * Milerate,
data = cot93, weights = wts.pace,
span = 0.05, degree = 1)

 Number of Observations:          5344
 Equivalent Number of Parameters: 44.4
 Residual Standard Error:         87.58
 Multiple R-squared:              0.64
 Residuals:
    min  1st Q median 3rd Q   max
 -370.7 -41.86  -5.22 32.84 547.1


[so, sigma_h^2  is 87.58174**2 or 7670.561181]

 M1[[i]]<-b+equivalent number of parameters

 V1[[i]]<-equivalent number of parameters


 cot.m.05$inference$s = Residual Standard Error

 cot.m.05$surface$enp = Equivalent Number of
                        Parameters


 for(i in seq(16)){
   s<-(i*.05)
   print(s)
   cot.m.05<-loess(Clhour ~ Ltdmile*Milerate, cot93,
weights=wts.pace, degree=1, span=s)
   RSE<-cot.m.05$inference$s
   enp<-cot.m.05$surface$enp
   sse<-sum((residuals(cot.m.05)*sqrt(wts.pace))**2)
   b<-(sse/7670.561181)-(sse/(RSE**2))
   M1[[i]]<-b+enp
   V1[[i]]<-enp
   }
```

APPENDIX C FORTRAN SUBROUTINES - MIXED MODEL

The author wishes to thank Dr. Richard H. Jones of the Department of Preventive Medicine and Biometrics, School of Medicine, University of Colorado, for providing the original version of a FORTRAN program for fitting a linear mixed model with polynomial growth curves. The original program has mainly been altered by the addition of the subroutines corr, random, and predict. These subroutines are detailed in this appendix. Other subroutines in the program may be found in *Longitudinal Data with Serial Correlation: A State-space Approach* by R.H. Jones [14]. The complete program used in this thesis may be obtained by contacting the author at kirchoff@umr.edu.

```
c NQ, number of random effects
c NOD, number of distinct elements in random effects covariance matrix
c      1 < NOD < NQ*(NQ+1)/2.  These elements are entered as the upper
c      triangular factorization of the covariance matrix.
c NR, maximum number of nonlinear parameters,
c     MODEL(1)+MODEL(2)+MODEL(3)+NOD
c NCV, number of covariates or grouping variables, constant for each
c      subject.
c NTVCV, number of covariates that vary within each subject's data
c NLP, number of linear parameters
c ND must be at least NLP+1
c NS, number of subjects
c NMAX, maximum number of observations
c
```

The corr subroutine calculates predicted values of the random effects, γ_i.

```
      subroutine corr(np,p,beta,like,rxy,randeff)
      parameter (nr=21,nlpmax=24,nd=nlpmax+1,ncvmax=5,nsmax=2000)
      parameter (nqmax=5,nodmax=15,nmax=10000,nomax=50,maxar=5,maxma=4)
      parameter (maxvxy=nomax+nlpmax+2+nqmax,nsnqmax=nqmax*nsmax)
      real y(nmax),t(nmax),cv(ncvmax,nsmax),p(np),beta(nlp)
      real u(nqmax,nqmax),tvcv(ncvmax,nmax)
      double precision x(nomax,nlpmax),z(nomax,nqmax)
      double precision xbeta(nomax),zb(nomax,nqmax),rdmeff(nqmax)
      double precision zu(nomax,nqmax),yy(nomax,nomax),pma(maxma)
      double precision ps(maxar),mse,xx(nd,nd),var
      double precision like,vxy(nomax,maxvxy),det,sse
      double precision rxy(nomax,maxvxy)
      double precision randeff(nsnqmax)
      complex*16 r(maxar),res(maxar),b1,b2,vv(nomax,nomax),varc
      integer nobs(nsmax),nxcv(ncvmax),in(nodmax,2),model(3),reml
      integer unit(nsmax)
      common /xrans/ nq,nlp,ns,xx,model,y,sse,nt,t,nobs,nod,in,
     +cv,nxcv,nx,ncv,ntvcv,tvcv,reml,unit
c
c  np      --> the number of nonlinear parameters
c  p       --> vector of nonlinear parameters
c  beta    --> estimates of fixed coefficients
c  like    <-- value of -2log likelihood
c  rxy     <-- total estimated error covariance matrix for subject i
c  randeff <-- predicted values of random coefficients
c
c  calculate roots of ar characteristic equation
c
      nar=model(1)
      nma=model(2)
      nlp1=nlp+1
      write(*,2000)(p(i),i=1,np)
 2000 format(' p(i) ',5g14.6)
      if(nar.gt.0)then
         do 1 i=1,nar
            ps(i)=p(i)
    1    continue
         call roots(nar,ps,r)
c
C  calculate denominators for covariance function
c
```

```
         do 100 j=1,nar
            res(j)=-2*real(r(j))
            do 90 k=1,nar
               if(k.ne.j)res(j)=res(j)*(r(k)-r(j))*(dconjg(r(k))+r(j))
   90       continue
  100    continue
         if(nma.gt.0)then
            call unma(nar,nma,p,pma)
         endif
      endif
      ncvx=nx
      ncvtot=ncv+ntvcv
      if (ncvtot.gt.0) then
         do 241 i=1,ncvtot
            ncvx=ncvx+nxcv(i)
  241    continue
      endif
      do 232 i=1,ncvx+1
         do 231 j=1,ncvx+1
            xx(i,j)=0.0
  231    continue
  232 continue
      if(nod.gt.0)then
         do 25 i=1,nq
            do 23 j=1,nq
               u(i,j)=0.0
   23       continue
   25    continue
         do 30 i=1,nod
            u(in(i,1),in(i,2))=p(nar+nma+model(3)+i)
   30    continue
         write(2,31)
   31         format(/' U matrix')
         do 33 i=1, nod
              write(2,32)in(i,1),in(i,2),u(in(i,1),in(i,2))
   32         format(1x,'u(',i1',',i1,')=',g15.8)
  33     continue
      endif
      nt=0
      det=0.0
c
c Start subject loop
c
      do 3000 ii=1,ns
```

```
c
c   Call for design matrices
c
         call xmatrix(ii,nx,t,nt,ncv,nxcv,cv,ntvcv,tvcv,nobs,x)
c  X matrix for subject 1
      If(ii.eq.1)then
         write(2,600)
 600     format(/' X matrix for subject 1 --corr')
         do 628 i=1,nobs(ii)
              write(2,627)(x(i,j),j=1,ncvx)
627           format(1x,7g11.4)
628           continue
      endif
         if (nod.gt.0) then
             call zmatrix(ii,t,nt,nobs,nq,z)
c
c  Z matrix for subject 1
      If(ii.eq.1.and.nod.gt.0)then
         write(2,26)
 26      format(/' Z matrix for subject 1 --corr')
         do 28 i=1,nobs(ii)
              write(2,27)(z(i,j),j=1,nq)
 27           format(1x,7g11.4)
 28           continue
      endif
c
c
c  calculate zu'
c
            do 50 i=1,nobs(ii)
               do 40 j=1,nq
                  zu(i,j)=0.0
                  do 35 k=j,nq
                     zu(i,j)=zu(i,j)+z(i,k)*u(j,k)
   35             continue
   40          continue
   50       continue
C
c  calculate error covariance matrix due to random effects
c
            do 80 i=1,nobs(ii)
               do 70 j=i,nobs(ii)
                  yy(i,j)=0.0
                  do 60 k=1,nq
```

```
                  yy(i,j)=yy(i,j)+zu(i,k)*zu(j,k)
   60             continue
                  yy(j,i)=yy(i,j)
   70          continue
   80       continue
         endif
C
c  calculate within subject error correlation matrix
c
         if (nar.gt.0) then
c
c First calculate process variance to normalize covariance matrix
c the mean square error, mse, estimates the within subject variance
c
            varc=(0.0d0,0.0d0)
            if(nma.eq.0)then
               do 1050 k=1,nar
                  varc=varc+1.0d0/res(k)
 1050          continue
            else
               do 1080 k=1,nar
                  b1=(1.0d0,0.0d0)
                  b2=(1.0d0,0.0d0)
                  isign=1
                  do 1070 l=1,nma
                     isign=-isign
                     b1=b1 + pma(l)*r(k)**l
                     b2=b2+isign*pma(l)*r(k)**l
 1070             continue
                  varc=varc+b1*b2/res(k)
 1080          continue
            endif
            var=varc
c
c Now calculate covariances and normalize to correlations
c
            do 120 i=1,nobs(ii)
               vxy(i,i)=1.0d0
               if(ii.eq.1)then
                  rxy(i,i)=vxy(i,i)
                  endif
               if(i+1.le.nobs(ii))then
                  do 110 j=i+1,nobs(ii)
                     time=t(j+nt)-t(i+nt)
```

```
                    vv(i,j)=(0.0d0,0.0d0)
                    if(nma.eq.0)then
                       do 105 k=1,nar
                          vv(i,j)=vv(i,j)+cdexp(r(k)*time)/res(k)
  105                  continue
                    else
                       do 108 k=1,nar
                          b1=(1.0d0,0.0d0)
                          b2=(1.0d0,0.0d0)
                          isign=1
                          do 107 l=1,nma
                             isign=-isign
                             b1=b1 + pma(l)*r(k)**l
                             b2=b2+isign*pma(l)*r(k)**l
  107                     continue
                          vv(i,j)=vv(i,j)+b1*b2*cdexp(r(k)*time)/res(k)
  108                  continue
                    endif
                    vxy(i,j)=vv(i,j)/var
                    vxy(j,i)=vxy(i,j)
                    if(ii.eq.1)then
                       rxy(i,j)=vxy(i,j)
                       rxy(j,i)=vxy(j,i)
                    endif
  110            continue
                 endif
                 if(model(3).eq.1)vxy(i,i)=vxy(i,i)+p(nar+nma+1)**2
  120         continue
           else
           do 125 i=1,nobs(ii)
              do 123 j=1,nobs(ii)
                 vxy(i,j)=0.0
  123         continue
              vxy(i,i)=1.0
  125      continue
        endif
C
c  calculate total error covariance matrix
c
        if(nod.gt.0)then
           do 140 i=1,nobs(ii)
              do 130 j=1,nobs(ii)
                 vxy(i,j)=vxy(i,j)+yy(i,j)
                  if(ii.eq.1)then
```

```
                        rxy(i,j)=vxy(i,j)
                endif
  130         continue
  140     continue
        endif
C
c  augment V by X
c
        do 410 i=1,nobs(ii)
           do 411 j=1,ncvx
              vxy(i,nobs(ii)+j)=x(i,j)
 411       continue
 410       continue

c  augment covariance and design matrices by observed data
c
        do 260 i=1,nobs(ii)
           vxy(i,nobs(ii)+nlp1)=y(nt+i)
  260 continue
c
c
        if(ii.eq.1.and.nod.gt.0) then
           write(2,270)
 270       format(/' VXY for subject 1 matrix from corr')
           do 271 i=1,nobs(ii)
              write(2,272)(vxy(i,j),j=1,nobs(ii)+nlp1+1)
 272          format(1x,7g11.4)
 271          continue
        endif
c
c  calculate estimates of random effects
c
        If(nod.gt.0)then
c  augment matrix with residuals
           do 250 i=1,nobs(ii)
              xbeta(i)=0.0
              do 251 j=nobs(ii)+1,nobs(ii)+ncvx
                 k=j-nobs(ii)
                 xbeta(i)=xbeta(i)+(vxy(i,j)*beta(k))
 251          continue
              vxy(i,nobs(ii)+nlp1+1)=y(nt+i)-xbeta(i)
 250       continue
c
c  calculate ZB and then augment matrix
```

```
c
         do 89 i=1,nobs(ii)
            do 91 j=1,nq
               zb(i,j)=0.0
               do 92 k=1,nq
                  zb(i,j)=zb(i,j)+(zu(i,k)*u(k,j))
 92            continue
               vxy(i,nobs(ii)+nlp1+1+j)=zb(i,j)
 91         continue
 89      continue
      call factor(vxy,nobs(ii),nomax,nobs(ii)+ncvx+2+nq,ier)
c
c  calculate estimates of random coefficients
c
      do 95 j=1,nq
         rdmeff(j)=0.0
         do 96 k=1,nobs(ii)
            rdmeff(j)=rdmeff(j)+(vxy(k,nobs(ii)+nlp1+1)*
     1vxy(k,nobs(ii)+nlp1+1+j))
 96      continue
 95   continue
c
c  store values of random effects
c
      do 400 i=1,nq
         randeff(i+(nq*(ii-1)))=rdmeff(i)
 400  continue
      endif
c
c  calculate and sum ln of |vi| for likelihood
c
 0    if(nod.gt.0) goto 98
        call factor(vxy,nobs(ii),nomax,nobs(ii)+ncvx+1,ier)
 98   do 280 i=1,nobs(ii)
         det=det+dlog(vxy(i,i)**2)
  280 continue
C
c  sum x'vx matrices for likelihood
c
      do 310 i=1,ncvx+1
          do 300 j=i,ncvx+1
             do 290 k=1,nobs(ii)
                xx(i,j)=xx(i,j)+vxy(k,nobs(ii)+i)*vxy(k,nobs(ii)+j)
  290        continue
```

```
  300     continue
  310 continue
      nt=nt+nobs(ii)
 3000  continue
c
c  calculate estimates of regression coefficients
c
      call factor(xx,ncvx,nd,ncvx+1,ier)
      sse=xx(ncvx+1,ncvx+1)
      do 311 i=1,ncvx
         sse=sse-xx(i,ncvx+1)**2
         if(reml.eq.1)then det=det+dlog(xx(i,i)**2)
  311 continue
      if(reml.eq.1)nt=nt-nlp
      mse=sse/nt
      like=nt*dlog(mse)+det+nt*2.837877067d0
      write(*,4000)like
 4000 format(' -2 ln likelihood= ',f12.3)
      return
      end
c
c
```

The random subroutine calculates standard errors of the predicted random coefficients by inverting the mixed model equations as given by Harville [13].

```
c
c
      subroutine random(np,p,xvx,randeff)
      parameter (nr=21,nlpmax=24,nd=nlpmax+1,ncvmax=5,nsmax=2000)
      parameter (nqmax=5,nodmax=15,nmax=10000,nomax=50,maxar=5,maxma=4)
      parameter (maxvxy=nomax+nlpmax+1,nqlpm=nqmax*nlpmax)
      parameter (ncnq=nomax+nlpmax+nqmax,nsnqmax=nsmax*nqmax)
      real y(nmax),t(nmax),cv(ncvmax,nsmax),p(np),tvcv(ncvmax,nmax)
      double precision z(nomax,nqmax),u(nqmax,nqmax)
      double precision b(nqmax,nqmax),x(nomax,nlpmax)
      double precision zu(nomax,nqmax),yy(nomax,nomax),pma(maxma)
      double precision ps(maxar),mse,xx(nd,nd),var
      double precision like,vxy(nomax,ncnq),det,sse
      double precision zb(nomax,nqmax),xvzb(nlpmax,nqmax)
      double precision bzvzb(nqmax,nqmax),mm2(nlpmax,nqmax)
      double precision mm4(nqmax,nqmax),stdv(nqlpm)
```

```
      double precision xvx(nlpmax,nlpmax),randeff(nsnqmax)
      complex*16 r(maxar),res(maxar),b1,b2,vv(nomax,nomax),varc
      integer nobs(nsmax),nxcv(ncvmax),in(nodmax,2),model(3),reml
      integer unit(nsmax)
      common /xrans/ nq,nlp,ns,xx,model,y,sse,nt,t,nobs,nod,in,
     +cv,nxcv,nx,ncv,ntvcv,tvcv,reml,unit
c
c
c  np       --> the number of nonlinear parameters
c  p        --> vector of nonlinear parameters
c  xvx      --> the matrix X'V(inv)X summed over all subjects
c  randeff <-- predicted values of random coefficients
c
c  calculate roots of ar characteristic equation
c
      nar=model(1)
      nma=model(2)
      nlp1=nlp+1
      write(*,2000)(p(i),i=1,np)
 2000 format(' p(i) ',5g14.6)
      if(nar.gt.0)then
         do 1 i=1,nar
            ps(i)=p(i)
    1    continue
         call roots(nar,ps,r)
c
C  calculate denominators for covariance function
c
         do 100 j=1,nar
            res(j)=-2*real(r(j))
            do 90 k=1,nar
               if(k.ne.j)res(j)=res(j)*(r(k)-r(j))*(dconjg(r(k))+r(j))
   90       continue
  100    continue
         if(nma.gt.0)then
            call unma(nar,nma,p,pma)
         endif
      endif
      ncvx=nx
      ncvtot=ncv+ntvcv
      if (ncvtot.gt.0) then
         do 241 i=1,ncvtot
            ncvx=ncvx+nxcv(i)
  241    continue
```

```
      endif
      if(nod.gt.0)then
         do 25 i=1,nq
            do 23 j=1,nq
               u(i,j)=0.0
   23       continue
   25    continue
         do 30 i=1,nod
            u(in(i,1),in(i,2))=p(nar+nma+model(3)+i)
   30    continue
      endif
      nts=nt
      nt=0
      mse=sse/nts
      rmse=sqrt(mse)
c
c  construct B matrix
c
      do 31 i=1,nq
         do 32 j=1,nq
            b(i,j)=0.0
            do 33 k=1,nq
               b(i,j)=b(i,j)+u(k,i)*u(k,j)
 33         continue
 32      continue
 31   continue
      write(2,34)
 34   format(2x,'Subject Unit   Est. random effect   Standard error')
c
c  Start subject loop
c
      do 3000 ii=1,ns
c
c   Call for design matrices
c
         call xmatrix(ii,nx,t,nt,ncv,nxcv,cv,ntvcv,tvcv,nobs,x)
c  X matrix for subject 1
      If(ii.eq.1)then
         write(2,600)
 600     format(/' X matrix for subject 1 --random')
         do 628 i=1,nobs(ii)
              write(2,627)(x(i,j),j=1,ncvx)
627           format(1x,7g11.4)
628           continue
```

```
      endif



         if (nod.gt.0) then
             call zmatrix(ii,t,nt,nobs,nq,z)
c  Z matrix for subject 1
      If(ii.eq.1.and.nod.gt.0)then
         write(2,26)
 26      format(/' Z matrix for subject 1 --random')
         do 28 i=1,nobs(ii)
              write(2,27)(z(i,j),j=1,nq)
 27           format(1x,7g11.4)
 28           continue
      endif
c
c
c   calculate zu'
c
            do 50 i=1,nobs(ii)
               do 40 j=1,nq
                  zu(i,j)=0.0
                  do 35 k=j,nq
                     zu(i,j)=zu(i,j)+z(i,k)*u(j,k)
   35             continue
   40          continue
   50       continue
C
c  calculate error covariance matrix due to random effects
c
            do 80 i=1,nobs(ii)
               do 70 j=i,nobs(ii)
                  yy(i,j)=0.0
                  do 60 k=1,nq
                     yy(i,j)=yy(i,j)+zu(i,k)*zu(j,k)
   60             continue
                  yy(j,i)=yy(i,j)
   70          continue
   80       continue
c
c  calculate ZB matrix
c
            do 91 i=1,nobs(ii)
               do 92 j=1,nq
```

```
                zb(i,j)=0.0
                do 93 k=1,nq
                   zb(i,j)=zb(i,j)+z(i,k)*b(k,j)
 93             continue
 92          continue
 91          continue
        endif
        if (nar.gt.0) then
c
c First calculate process variance to normalize covariance matrix
c the mean square error, mse, estimates the within subject variance
c
          varc=(0.0d0,0.0d0)
          if(nma.eq.0)then
             do 1050 k=1,nar
                varc=varc+1.0d0/res(k)
 1050        continue
          else
             do 1080 k=1,nar
                b1=(1.0d0,0.0d0)
                b2=(1.0d0,0.0d0)
                isign=1
                do 1070 l=1,nma
                   isign=-isign
                   b1=b1 + pma(l)*r(k)**l
                   b2=b2+isign*pma(l)*r(k)**l
 1070           continue
                varc=varc+b1*b2/res(k)
 1080        continue
          endif
          var=varc
c
c Now calculate covariances and normalize to correlations
c
          do 120 i=1,nobs(ii)
             vxy(i,i)=1.0d0
             if(i+1.le.nobs(ii))then
                do 110 j=i+1,nobs(ii)
                   time=t(j+nt)-t(i+nt)
                   vv(i,j)=(0.0d0,0.0d0)
                   if(nma.eq.0)then
                      do 105 k=1,nar
                         vv(i,j)=vv(i,j)+cdexp(r(k)*time)/res(k)
 105                  continue
```

```
                    else
                       do 108 k=1,nar
                          b1=(1.0d0,0.0d0)
                          b2=(1.0d0,0.0d0)
                          isign=1
                          do 107 l=1,nma
                             isign=-isign
                             b1=b1 + pma(l)*r(k)**l
                             b2=b2+isign*pma(l)*r(k)**l
  107                     continue
                          vv(i,j)=vv(i,j)+b1*b2*cdexp(r(k)*time)/res(k)
  108                  continue
                    endif
                    vxy(i,j)=vv(i,j)/var
                    vxy(j,i)=vxy(i,j)
  110            continue
              endif
              if(model(3).eq.1)vxy(i,i)=vxy(i,i)+p(nar+nma+1)**2
  120       continue
         else
         do 125 i=1,nobs(ii)
            do 123 j=1,nobs(ii)
               vxy(i,j)=0.0
  123       continue
            vxy(i,i)=1.0
  125    continue
      endif
C
c  calculate total error covariance matrix
c
      if(nod.gt.0)then
         do 140 i=1,nobs(ii)
            do 130 j=1,nobs(ii)
               vxy(i,j)=vxy(i,j)+yy(i,j)
  130       continue
  140    continue
      endif
c
c  augment V by X and ZB
c
      do 400 i=1,nobs(ii)
         do 401 j=1,ncvx
            vxy(i,nobs(ii)+j)=x(i,j)
 401     continue
```

```
 400  continue
      do 410 i=1,nobs(ii)
         do 411 j=1,nq
            vxy(i,nobs(ii)+ncvx+j)=zb(i,j)
 411     continue
 410  continue
c
c  factor V
c
      call factor(vxy,nobs(ii),nomax,nobs(ii)+ncvx+nq,ier)
c
c  calculate X'V(inv)ZB, BZ'V(inv)ZB
c
      do 420 i=1,ncvx
         do 421 j=1,nq
            xvzb(i,j)=0.0
            do 422 k=1,nobs(ii)
               xvzb(i,j)=xvzb(i,j)+vxy(k,nobs(ii)+i)*
     1vxy(k,nobs(ii)+ncvx+j)
 422        continue
 421     continue
 420  continue
      do 430 i=1,nq
         do 431 j=1,nq
            bzvzb(i,j)=0.0
            do 432 k=1,nobs(ii)
               bzvzb(i,j)=bzvzb(i,j)+vxy(k,nobs(ii)+ncvx+i)*
     1vxy(k,nobs(ii)+ncvx+j)
 432        continue
 431     continue
 430  continue
c
c  calculate upper right quadrant of inverted coefficient matrix
c  of mixed model equations
c
      do 440 i=1,ncvx
         do 441 j=1,nq
            mm2(i,j)=0.0
            do 442 k=1,ncvx
               mm2(i,j)=mm2(i,j)+(-xvx(i,k)*xvzb(k,j))
 442        continue
 441     continue
 440  continue
c
```

```
c  calculate lower right quadrant of inverted coefficient matrix
c  of mixed model equations
c
      do 450 i=1,nq
         do 451 j=1,nq
            mm4(i,j)=b(i,j)-bzvzb(i,j)
            do 452 k=1,ncvx
               mm4(i,j)=mm4(i,j)-(xvzb(k,i)*mm2(k,j))
 452        continue
 451     continue
 450  continue
c
c  calculate standard errors of fixed and random effects
c
      do 500 i=1,ncvx
         do 501 j=1,ncvx
            stdv(i)=rmse*sqrt(xvx(i,i))
 501     continue
 500  continue
      do 510 i=1,nq
         do 511 j=1,nq
            stdv(ncvx+i)=rmse*sqrt(mm4(i,i))
 511     continue
 510  continue
c
c  write estimates and standard errors of random effects
c
      do 530 i=1,nq
         write(2,540)ii,unit(ii),randeff(i+(nq*(ii-1))),stdv(ncvx+i)
 540     format(2x,i5,2x,i7,7x,7g11.4,7x,7g11.4)
 530  continue
       nt=nt+nobs(ii)
 3000  continue
       return
       end
c
c
```

The predict subroutine calculates predicted responses and their estimated standard errors at the observed time points.

```
c
c
      subroutine predict(np,p,xvx,beta,randeff)
      parameter (nr=21,nlpmax=24,nd=nlpmax+1,ncvmax=5,nsmax=2000)
      parameter (nqmax=5,nodmax=15,nmax=10000,nomax=50,maxar=5,maxma=4)
      parameter (maxvxy=nomax+nlpmax+1,nsnqmax=nsmax*nqmax)
      parameter (ncnq=nomax+nlpmax+nqmax,nqlpm=nqmax*nlpmax)
      real y(nmax),t(nmax),cv(ncvmax,nsmax),p(np),beta(nlpmax)
      real tvcv(ncvmax,nmax)
      double precision z(nomax,nqmax),u(nqmax,nqmax)
      double precision b(nqmax,nqmax),x(nomax,nlpmax)
      double precision zu(nomax,nqmax),yy(nomax,nomax),pma(maxma)
      double precision ps(maxar),mse,xx(nd,nd),var
      double precision like,vxy(nomax,ncnq),det,sse
      double precision zb(nomax,nqmax),xvzb(nlpmax,nqmax)
      double precision bzvzb(nqmax,nqmax),mm2(nlpmax,nqmax)
      double precision mm4(nqmax,nqmax),stdv(nqlpm)
      double precision xvx(nlpmax,nlpmax)
      double precision randeff(nsnqmax)
      double precision tau(nomax),vr(nomax),xmz(nomax)
      double precision zmz(nomax),stdep(nomax)
      complex*16 r(maxar),res(maxar),b1,b2,vv(nomax,nomax),varc
      integer nobs(nsmax),nxcv(ncvmax),in(nodmax,2),model(3),reml
      integer unit(nsmax)
      common /xrans/ nq,nlp,ns,xx,model,y,sse,nt,t,nobs,nod,in,
     +cv,nxcv,nx,ncv,ntvcv,tvcv,reml,unit
c
c
c  np      --> the number of nonlinear parameters
c  p       --> vector of nonlinear parameters
c  xvx     --> the matrix X'V(inv)X summed over all subjects
c  beta    --> estimated of fixed coefficients
c  randeff <-- predicted values of random coefficients
c
c
c  calculate roots of ar characteristic equation
c
      nar=model(1)
      nma=model(2)
```

```
      nlp1=nlp+1
      write(*,2000)(p(i),i=1,np)
 2000 format(' p(i) ',5g14.6)
      if(nar.gt.0)then
         do 1 i=1,nar
            ps(i)=p(i)
    1    continue
         call roots(nar,ps,r)
c
C  calculate denominators for covariance function
c
         do 100 j=1,nar
            res(j)=-2*real(r(j))
            do 90 k=1,nar
               if(k.ne.j)res(j)=res(j)*(r(k)-r(j))*(dconjg(r(k))+r(j))
   90       continue
  100    continue
         if(nma.gt.0)then
            call unma(nar,nma,p,pma)
         endif
      endif
      ncvx=nx
      ncvtot=ncv+ntvcv
      if (ncvtot.gt.0) then
         do 241 i=1,ncvtot
            ncvx=ncvx+nxcv(i)
  241    continue
      endif
      if(nod.gt.0)then
         do 25 i=1,nq
            do 23 j=1,nq
               u(i,j)=0.0
   23       continue
   25    continue
         do 30 i=1,nod
            u(in(i,1),in(i,2))=p(nar+nma+model(3)+i)
   30    continue
      endif
      nts=nt
      nt=0
      mse=sse/nts
      rmse=sqrt(mse)
      write(2,34)
 34   format(1x,'Subject      Unit    Time       Observed
```

```
     1Predicted  Std error')
c
c  construct B matrix
c
      if(nod.gt.0)then
      do 31 i=1,nq
         do 32 j=1,nq
            b(i,j)=0.0
            do 33 k=1,nq
               b(i,j)=b(i,j)+u(k,i)*u(k,j)
 33         continue
 32      continue
 31   continue
      endif
c
c  Start subject loop
c
      do 3000 ii=1,ns
c
c   Call for design matrices
c
         call xmatrix(ii,nx,t,nt,ncv,nxcv,cv,ntvcv,tvcv,nobs,x)
c   Go to prediction if there are no random effects
c
       If(nod.eq.0)then
          goto 600
       endif
c
c  construct Z matrix
c
         if(nod.gt.0)then
             call  zmatrix(ii,t,nt,nobs,nq,z)
c
c  calculate zu'
c
           do 50 i=1,nobs(ii)
              do 40 j=1,nq
                 zu(i,j)=0.0
                 do 35 k=j,nq
                    zu(i,j)=zu(i,j)+z(i,k)*u(j,k)
   35            continue
   40         continue
   50       continue
C
```

```
c  calculate error covariance matrix due to random effects
c
            do 80 i=1,nobs(ii)
               do 70 j=i,nobs(ii)
                  yy(i,j)=0.0
                  do 60 k=1,nq
                     yy(i,j)=yy(i,j)+zu(i,k)*zu(j,k)
   60             continue
                  yy(j,i)=yy(i,j)
   70          continue
   80       continue
c
c  calculate ZB matrix
c
            do 91 i=1,nobs(ii)
               do 92 j=1,nq
                  zb(i,j)=0.0
                  do 93 k=1,nq
                     zb(i,j)=zb(i,j)+z(i,k)*b(k,j)
 93               continue
 92            continue
 91            continue
         endif
         if (nar.gt.0) then
c
c First calculate process variance to normalize covariance matrix
c the mean square error, mse, estimates the within subject variance
c
            varc=(0.0d0,0.0d0)
            if(nma.eq.0)then
               do 1050 k=1,nar
                  varc=varc+1.0d0/res(k)
 1050          continue
            else
               do 1080 k=1,nar
                  b1=(1.0d0,0.0d0)
                  b2=(1.0d0,0.0d0)
                  isign=1
                  do 1070 l=1,nma
                     isign=-isign
                     b1=b1 + pma(l)*r(k)**l
                     b2=b2+isign*pma(l)*r(k)**l
 1070             continue
                  varc=varc+b1*b2/res(k)
```

```
 1080          continue
            endif
            var=varc
c
c Now calculate covariances and normalize to correlations
c
            do 120 i=1,nobs(ii)
               vxy(i,i)=1.0d0
               if(i+1.le.nobs(ii))then
                  do 110 j=i+1,nobs(ii)
                     time=t(j+nt)-t(i+nt)
                     vv(i,j)=(0.0d0,0.0d0)
                     if(nma.eq.0)then
                        do 105 k=1,nar
                           vv(i,j)=vv(i,j)+cdexp(r(k)*time)/res(k)
 105                    continue
                     else
                        do 108 k=1,nar
                           b1=(1.0d0,0.0d0)
                           b2=(1.0d0,0.0d0)
                           isign=1
                           do 107 l=1,nma
                              isign=-isign
                              b1=b1 + pma(l)*r(k)**l
                              b2=b2+isign*pma(l)*r(k)**l
 107                       continue
                           vv(i,j)=vv(i,j)+b1*b2*cdexp(r(k)*time)/res(k)
 108                    continue
                     endif
                     vxy(i,j)=vv(i,j)/var
                     vxy(j,i)=vxy(i,j)
 110              continue
               endif
               if(model(3).eq.1)vxy(i,i)=vxy(i,i)+p(nar+nma+1)**2
 120        continue
         else
         do 125 i=1,nobs(ii)
            do 123 j=1,nobs(ii)
               vxy(i,j)=0.0
 123        continue
            vxy(i,i)=1.0
 125     continue
      endif
C
```

```
c  calculate total error covariance matrix
c
      if(nod.gt.0)then
         do 140 i=1,nobs(ii)
            do 130 j=1,nobs(ii)
               vxy(i,j)=vxy(i,j)+yy(i,j)
  130       continue
  140    continue
      endif
c
c  augment V by X and ZB
c
      do 400 i=1,nobs(ii)
         do 401 j=1,ncvx
            vxy(i,nobs(ii)+j)=x(i,j)
 401     continue
 400  continue
      do 410 i=1,nobs(ii)
         do 411 j=1,nq
            vxy(i,nobs(ii)+ncvx+j)=zb(i,j)
 411     continue
 410  continue
c
c  factor V
c
      call factor(vxy,nobs(ii),nomax,nobs(ii)+ncvx+nq,ier)
c
c  calculate X'V(inv)ZB, BZ'V(inv)ZB
c
      do 420 i=1,ncvx
         do 421 j=1,nq
            xvzb(i,j)=0.0
            do 422 k=1,nobs(ii)
               xvzb(i,j)=xvzb(i,j)+vxy(k,nobs(ii)+i)*
     1vxy(k,nobs(ii)+ncvx+j)
 422        continue
 421     continue
 420  continue
      do 430 i=1,nq
         do 431 j=1,nq
            bzvzb(i,j)=0.0
            do 432 k=1,nobs(ii)
               bzvzb(i,j)=bzvzb(i,j)+vxy(k,nobs(ii)+ncvx+i)*
     1vxy(k,nobs(ii)+ncvx+j)
```

```
 432        continue
 431     continue
 430  continue
c
c  calculate upper right quadrant of inverted coefficient matrix
c  of mixed model equations
c
      do 440 i=1,ncvx
         do 441 j=1,nq
            mm2(i,j)=0.0
            do 442 k=1,ncvx
               mm2(i,j)=mm2(i,j)+(-xvx(i,k)*xvzb(k,j))
 442        continue
 441     continue
 440  continue
c
c  calculate lower right quadrant of inverted coefficient matrix
c  of mixed model equations
c
      do 450 i=1,nq
         do 451 j=1,nq
            mm4(i,j)=b(i,j)-bzvzb(i,j)
            do 452 k=1,ncvx
               mm4(i,j)=mm4(i,j)-(xvzb(k,i)*mm2(k,j))
 452        continue
 451     continue
 450  continue
c
c  calculate standard errors of fixed and random effects
c
      do 500 i=1,ncvx
         do 501 j=1,ncvx
            stdv(i)=rmse*sqrt(xvx(i,i))
 501     continue
 500  continue
      do 510 i=1,nq
         do 511 j=1,nq
            stdv(ncvx+i)=rmse*sqrt(mm4(i,i))
 511     continue
 510  continue
c
c  calculate predictions and std errors of prediction
c
 600  do 610 i=1,nobs(ii)
```

```
        tau(i)=0
        do 611 k=1,ncvx
           tau(i)=tau(i)+x(i,k)*beta(k)
611     continue
610  continue
     do 620 i=1,nobs(ii)
        vr(i)=0.0
        do 621 j=1,ncvx
           do 622 k=1,ncvx
              vr(i)=vr(i)+(x(i,j)*xvx(j,k)*x(i,k))
622        continue
621     continue
620  continue
     if(nod.eq.0)then
        do 625 i=1,nobs(ii)
           stdep(i)=rmse*sqrt(vr(i))
625     continue
        do 627 i=1,nobs(ii)
           write(2,630) ii,unit(ii),x(i,2),y(nt+i),tau(i),stdep(i)
630        format(i4,8x,i7,4x,7g11.4,7g11.4,7g11.4,7g11.4)
627     continue
        goto 700
     else
        do 640 i=1,nobs(ii)
           do 641 k=1,nq
              tau(i)=tau(i)+z(i,k)*randeff(k+(nq*(ii-1)))
641        continue
640     continue
        do 650 i=1,nobs(ii)
           xmz(i)=0.0
           do 651 j=1,ncvx
              do 652 k=1,nq
                 xmz(i)=xmz(i)+2*(x(i,j)*mm2(j,k)*z(i,k))
652           continue
651        continue
650     continue
        do 660 i=1,nobs(ii)
           zmz(i)=0.0
           do 661 j=1,nq
              do 662 k=1,nq
                 zmz(i)=zmz(i)+z(i,j)*mm4(j,k)*z(i,k)
662           continue
661        continue
660     continue
```

```
        do 670 i=1,nobs(ii)
           vr(i)=vr(i)+xmz(i)+zmz(i)
           stdep(i)=rmse*sqrt(vr(i))
670     continue
        do 672 i=1,nobs(ii)
           write(2,675) ii,unit(ii),x(i,2),y(nt+i),tau(i),stdep(i)
675        format(i4,8x,i7,4x,7g11.4,7g11.4,7g11.4,7g11.4)
672     continue
     endif
700  nt=nt+nobs(ii)
3000  continue
      return
      end
```

BIBLIOGRAPHY

[1] Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, 85, 749-59.

[2] Chambers, J.M. and Hastie, T.J. (1992). Chapter 8, Local regression models. *Statistical Models in S.* Wadsworth and Brooks, Pacific Grove, California.

[3] Clark, R.M. (1980). Calibration, cross-validation, and carbon 14 II. *Journal of the Royal Statistical Society, Series A*, 143, 177-94.

[4] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

[5] Cleveland, W.S., and Devlin, S.J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.

[6] Cleveland, W.S., Devlin, S.J, and Grosse, E. (1988). Regression by local fitting. Methods, properties, and computational algorithms. *Journal of Econometrics*, 37, 87-114.

[7] Diggle, J., Liang, K., and Zeger, S. (1994). Chapter 3, Exploring longitudinal data. *Analysis of Longitudinal Data.* Oxford University Press, Oxford, England.

[8] Efron, B. and Tibshirani, R.J. (1992). Chapter 6, The bootstrap estimate of standard error. *An Introduction to the Bootstrap.* Chapman and Hall, London, England.

[9] Gasser, T., Müller, H.G., Köhler, W., Molinari, L., and Prader, A. (1984). Nonparametric regression analysis of growth curves. *Annals of Statistics*, 12, 210-29.

[10] Grizzle, J. and Allen, D. (1969). Analysis of growth and dose response curves. *Biometrics*, 25, 357-82.

[11] Hart, J.D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society*, **B**, 53, 173-187.

[12] Hart, J.D., and Wehrly, T.E. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, 81, 1080-88.

[13] Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.

[14] Jones, R. (1993). *Longitudinal Data with Serial Correlation: A State-space Approach.* Chapman and Hall, London, England.

[15] Jones, R. and Ackerson, L. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika*, 77, 721-31.

[16] Jones, R. and Boadi-Boateng, F. (1991). Unequally spaced longitudinal data with ar(1) serial correlation. *Biometrics*, 47, 161-175.

[17] Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

[18] Müller, H. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *American Statistical Association*, 82, 231-238.

[19] Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika,*, 51, 313-26.

[20] Priestley, M.B. and Chao, M.T. (1972). Nonparametric function fitting. *Journal of the Royal Statistical Society, Series B*, 34, 385-92.

[21] Rao, C.R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, 46, 49-58.

[22] Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of grwoth curves. *Biometrika*, 52, 447-58.

[23] Rao, C.R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* 31, 545-554.

[24] Rao, C.R. (1987). Prediction of future observations in growth curve models. *Statistical Science* 2, 4, 434-471.

[25] Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association* 79, 406-414.